MULTIVARIATE POISSON HIDDEN MARKOV MODELS FOR ANALYSIS OF SPATIAL COUNTS

A Thesis Submitted to the Faculty of Graduate Studies and Research in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

in the Department of Mathematics and Statistics
University of Saskatchewan, Saskatoon,
SK, Canada

by

Chandima Piyadharshani Karunanayake

PERMISSION TO USE

The author has agreed that the libraries of this University may provide the thesis freely available for inspection. Moreover, the author has agreed that permission for copying of the thesis in any manner, entirely or in part, for scholarly purposes may be granted by the Professor or Professors who supervised my thesis work or in their absence, by the Head of the Department of Mathematics and Statistics or the Dean of the College in which the thesis work was done. It is understood that any copying or publication or use of the thesis or parts thereof for finanancial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to the author and to the University of Saskatchewan in any scholarly use which may be made of any material in this thesis.

Requests for permission to copy or to make other use of any material in the thesis should be addressed to:

Head

Department of Mathematics and Statistics

University of Saskatchewan

106, Wiggins Road

Saskatoon, Saskatchewan

Canada, S7N 5E6

ABSTRACT

Multivariate count data are found in a variety of fields. For modeling such data, one may consider the multivariate Poisson distribution. Overdispersion is a problem when modeling the data with the multivariate Poisson distribution. Therefore, in this thesis we propose a new multivariate Poisson hidden Markov model based on the extension of independent multivariate Poisson finite mixture models, as a solution to this problem. This model, which can take into account the spatial nature of weed counts, is applied to weed species counts in an agricultural field. The distribution of counts depends on the underlying sequence of states, which are unobserved or hidden. These hidden states represent the regions where weed counts are relatively homogeneous. Analysis of these data involves the estimation of the number of hidden states, Poisson means and covariances. Parameter estimation is done using a modified EM algorithm for maximum likelihood estimation.

We extend the univariate Markov-dependent Poisson finite mixture model to the multivariate Poisson case (bivariate and trivariate) to model counts of two or three species. Also, we contribute to the hidden Markov model research area by developing Splus/R codes for the analysis of the multivariate Poisson hidden Markov model. Splus/R codes are written for the estimation of multivariate Poisson hidden Markov model using the EM algorithm and the forward-backward procedure and the bootstrap estimation of standard errors. The estimated parameters are used to calculate the goodness of fit measures of the models.

Results suggest that the multivariate Poisson hidden Markov model, with five states and an independent covariance structure, gives a reasonable fit to this dataset. Since this model deals with overdispersion and spatial information, it will help to get an insight about weed distribution for herbicide applications. This model may lead researchers to find other factors such as soil moisture, fertilizer level, etc., to determine the states, which govern the distribution of the weed counts.

Keywords: Multivariate Poisson distribution, multivariate Poisson hidden Markov model, Weed species counts, EM algorithm.

ACKNOWLEDGEMENT

First I would like to acknowledge and express my sincere thanks and gratitude to my supervisor Dr. William H. Laverty for his availability, continual guidance, valuable suggestions and encouragement throughout the course of study.

Next, I would like to thank the members of my advisory committee, Prof. R. Srinivasan, Prof. C. E. Soteros, Prof. M.J. Miket and Prof. I.W. Kelly for their valuable suggestions and advice in many aspects of my thesis completion. I am also grateful for comments and suggestions from my external examiner, Prof. Peter MacDonald.

My special thanks to Dr. Dimitris Karlis, Athens University of Economics, Athens, Greece for his valuable advice and help me solve the problems I had with multivariate Poisson distributions.

I am very grateful for the funding provided by College of Graduate Studies and Research and Depaerment of Mathematics and Statistics. Without their support and resources, it is impossible to complete this thesis.

I would also like to thank Ms. Jessica Antonio of the Department of English, University of Saskatchewan for proof reading this thesis.

Finally, my heartfelt thanks go to my dear parents, and especially my husband, Sumith Priyashantha, who always wished and encouraged me to successfully complete my study program in Canada.

DEDICATION

This thesis is dedicated to my loving parents, Prof. Marcus Marcy Karunanayake and Mrs. Sumana Piyaseeli Karunanayake, and my dearest husband, Kahanda Rathmalapage Sumith Priyashantha, who always gave me encouragement for the success in my academic career.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ACRONYMS

AIC	Akaike Information Criterion
BIC	Bayseian Information Criterion
EM	Expectation- Maximization
GIS	Geographic Information System
GLM	Generalized Linear Model
HMM	Hidden Markov Model
HMRF	Hidden Markov Random Field
LRT	Likelihood Ratio Test
MFM	Multivariate Finite Mixture
ML	Maxiumum Likelihood
MRF	Markov Random Field
SEM	Stochastic Expectation- Maximization

CHAPTER 1

GENERAL INTRODUCTION

1.1 Introduction

The analysis of multivariate count data (e.g. weed counts for different species in a field) that are overdispersed relative to the Poisson distribution (i.e. variance > mean) has recently received considerable attention (Karlis and Meligkotsidou, 2006; Chib and Winkelmann, 2001). Such data might arise in an agricultural field study where overdispersion is caused by the individual variability of experimental units, soil types or fertilizer levels. Therefore, these data (e.g. weed counts) are not homogenous within the field. The Poisson mixture model is a flexible alternative model which can represent the inhomogeneous population. Finite Poisson mixtures are very popular for clustering since they lead to a simple and natural interpretation, as models describing a population consisting of a finite number of subpopulations.

These types of count data can be modelled using model-based clustering methods, such as multivariate Poisson finite mixture models (or independent finite mixture models) and multivariate Poisson hidden Markov models (or Markov-dependent finite mixture models). It is assumed that the counts follow independent Poisson distributions

conditional on rates, which are generated from an independent mixing distribution for finite mixture models. The counts for multivariate Poisson hidden Markov models are assumed to follow independent Poisson distributions, conditional on rates with Markov dependence. Finite mixture models can be particularly attractive because they provide plausible explanations for variation in the data (Leroux and Puterman, 1992).

1.2 Literature review

1.2.1 Introduction to finite mixture models

The main question here is determining the structure of clustered data when no information other than the observed values is available. Finite mixture models have been proposed for quite sometime as a basis for studying the clustered data (Symons, 1981; McLachan, 1982; McLachlan et al., 1988). In this approach, the data are viewed as coming from a mixture of probability distributions, each representing a different cluster. Recently, finite mixture model analysis have been used in several practical applications: character recognition (Murtagh and Raftery, 1984); tissue segmentation (Banfield and Raftery, 1993); minefield and seismic fault detection (Dasgupta and Raftery, 1998); identification of textile flaws from images (Campbell et al., 1997); and classification of astronomical data (Celeux et al., 1995). Most of these examples are based on Gaussian finite mixture models. There are some examples of Poisson finite mixtures. Leroux and Puterman (1992) describe a univariate Poisson finite mixture model for fetal movement data. The clustering of cases of a rare disease (sudden infant death syndrome), on the basis of the number of cases observed for various counties in

North Carolina, is modelled by Symons et al. (1983) using a mixture of two Poisson distributions, which describe the two groups of high and low risk counties. Very recently, a multivariate Poisson finite mixture model was used for a marketing application (Brijs et al., 2004). Brijs describes a multivariate Poisson finite mixture model for clustering supermarket shoppers based on their purchase frequency in a set of product categories.

In this thesis, the multivariate Poisson finite mixture model is applied for the first time to weed species counts in an agricultural field. Also, we developed a multivariate Poisson hidden Markov model and applied it to analyze the weed species data. The goodness of fit measure of the model is also evaluated (Chapter 7). Details about multivariate Poisson finite mixture models and multivariate Poisson hidden Markov models are given in Chapter 5. The history of hidden Markov models is presented in the next section.

1.2.2 History of hidden Markov models

Hidden Markov Models (HMMs) are statistical models that are widely used in many areas of probabilistic modeling. These models have received increasing attention (Rabiner and Juang, 1986, 1991 and Rabiner, 1989), partially because of their mathematical properties (they are rich in mathematical structure), but mostly because of their applications to many important areas in scientific research.

Hidden Markov Models have been found to be extremely useful for modeling stock market behavior. For example, the quarterly change in the exchange rate of the dollar can be modelled as an HMM with two states, which are unobservable and correspond to the up and down changes in exchange rate (Engel and Hamilton, 1990). HMM is also used in the area of speech recognition. Juang and Rabiner (1991) and Rabiner (1989) described how one could design a distinct hidden Markov model for each word in one's vocabulary, in order to envision the physical meaning of the model states as distinct sounds (e.g. Phonemes, syllables). A hidden Markov model for ecology was introduced by Baum and Eagon (1967). Later, they introduced a procedure for the maximum likelihood estimation of the HMM parameters for the general case where the observed sequence is a sequence of random variables with log-concave densities (Baum et al., 1970). In molecular biology, hidden Markov models are used to allow for unknown rates of evolution at different sites in a molecular sequence (Felsenstein and Churchill, 1996). Similarly, in climatology, the occurrence or nonoccurrence of rainfall at different sites can be modelled as an HMM where the climate states are unobservable, accounting for different distributions of rainfall over the sites (Zucchini et al., 1991).

The distinction between non-hidden Markov models and hidden Markov models is based on whether the output of the model is the actual state sequence of the Markov model, or if the output is an observation sequence generated from the state sequence. For hidden Markov models, the output is not the state sequence, but observations that are probabilistic function of the states. Thus in hidden Markov models, it extends the

concept of Markov models to include the case where the observation is a probabilistic function of the states.

The concept of hidden Markov Model has been the object of considerable study since the basic theory of hidden Markov models was initially introduced and studied during the late 1960's and early 1970's by Baum and his colleagues (Baum et al., 1966, 1967 and 1970). The primary concern in the hidden Markov modeling technique is the estimation of the model parameters from the observed sequences. One method of estimating the parameters of the hidden Markov models is to use the well-known Baum-Welch re-estimation method (Baum and Petrie, 1966). Baum and Eagon first proposed the algorithm in 1967 for the estimation problem of hidden Markov models with discrete observation densities. Baum and others (1970) later extended this algorithm to continuous density hidden Markov models with some limitations.

1.2.3 Hidden Markov model and hidden Markov random field model

Hidden Markov models are well known models in modeling the unknown state sequence given the observation sequence. As mentioned in the previous section, this has been successfully applied in the fields of speech recognition, biological modeling (protein sequences and DNA sequences) and many other fields. The hidden Markov models presented in section 1.2.2 are one-dimensional models, and they cannot take spatial dependencies into account. To overcome this drawback, Markov random fields and hidden Markov random fields (HMRF) can be used in more than one dimension when considering the spatial dependencies. For example, when the state space or

locations have two coordinates, that state space can be considered as a two-dimensional nearest-neighbor Markov random field. These Markov random fields have been extensively applied in the field of image processing (Fjϕrtoft et al., 2003; Pieczynski et al., 2002; Zhang et al., 2001; Fjϕrtoft et al., 2001; Aas et al., 1999).

In each case, there is a set of quantities, x, representing some unobservable phenomenon, and a supplementary set of observables, y. In general, y is a distorted version of x. For example, in the context of speech recognition, x represents a time sequence of configurations of an individual's vocal tract; the y represents the corresponding time sequence of projected sounds. Here, the Markovian assumption would be that the elements of x come form a realization of a Markov chain. In the context of image analysis, x represents the true scene, in terms of the true pixellated colouring, and y denotes the corresponding observed image. Here, the Markovian assumption would be that the elements of x would be assumed to come from a Markov random field. The elements of x are indexed by a set, S, of sites, usually representing time-points or discrete points in space (Archer & Titterington, 2002).

There is a very close relationship between Markov random fields and Markov chains. Estimation of Markov random field prior parameters can be done using Markov chain Monte Carlo Maximum likelihood estimation (Descombes et al., 1999). It is also demonstrated that a 2-D Markov random field can be easily transformed into a one-dimensional Markov chain (Fjϕrtoft et al., 2003). Fjϕrtoft (2003) explains that in image analysis, hidden Markov random field (HMRF) models are often used to impose spatial

regularity constraints on the underlying classes of an observed image, which allow Bayesian optimization of the classification. However, the computing time is often prohibitive with this approach. A substantially quicker alternative is to use a hidden Markov model (HMM), which can be adapted to two-dimensional analysis through different types of scanning methods (e.g. Line Scan, Hilbert-Peano scan etc.). Markov random field models can only be used for small neighbourhoods in the image, due to the computational complexity and the modeling problems posed by large neighbourhoods (Aas et al., 1999). Leroux and Puterman (1992) used maximum–penalized likelihood estimation to estimate the independent and the Markov-dependent mixture model parameters. In their analysis, they focus on the use of Poisson mixture models assuming independent observations and Markov-dependent models (or hidden Markov Models) for a set of univariate fetal movement counts. Extending this idea, for a set of multivariate Poisson counts, a novel multivariate Poisson hidden Markov model (Markov-dependent multivariate Poisson finite mixture model) is introduced. These counts can be considered as a stochastic process, generated by a Markov chain whose state sequence cannot be observed directly but which can be indirectly estimated through observations. Zhang et al. (2001) described that the finite mixture model is a degenerate version of the hidden Markov random field model. Fjϕrtoft (2003) explained that the classification accuracy of hidden Markov random fields and hidden Markov models were not differing very much. Hidden Markov models are much faster than the ones based on the Markov random fields (Fjϕrtoft et al., 2003). The advantage of hidden Markov models compared to the Markov random field models is the ability to combine

the simplicity of local modeling with the strength of global dependence by considering one-dimensional neighbourhoods (Aas et al., 1999).

1.3 Outline of the thesis

Chapter 2 gives a review of the Markov process, and then gives examples of hidden Markov models to clarify and present the general definition of the HMM and the HMRF. Chapter 3 is about the prediction, the state identification and the estimation problem, the solution of the HMM for a univariate case. The question of interest in this thesis is presented in Chapter 4. Details about calculating multivariate Poisson probabilities, multivariate Poisson finite mixture models and multivariate Poisson hidden Markov models are discussed in Chapter 5. We extended the univariate Markov-dependent Poisson mixture model to a multivariate Poisson case (bivariate and trivariate). Also, we contributed to the hidden Markov model research area by developing Splus/R codes for the analysis of the multivariate Poisson hidden Markov Model. Splus/R codes are written to estimate the multivariate Poisson hidden Markov Model using the EM algorithm and the forward-backward procedure and the bootstrap estimation of standard errors. Results are presented in Chapter 6. The properties of the finite mixture models and several applications are presented in Chapter 7. The Computational efficiency of the models is discussed in Chapter 8. The discussion, conclusion and the areas of further research are presented in Chapter 9.

CHAPTER 2

HIDDEN MARKOV MODELS (HMM's) AND HIDDEN MARKOV RANDOM FIELDS(HMRF's)

2.1 Discrete time finite state Markov chain

Let $\{S_t, t = 0,1,2,....\}$ be a sequence of integer valued random variables that can assume only an integer value $\{1,2,....,K\}$. Then $\{S_t, t = 0,1,2,....\}$ is a K state Markov chain if the probability that S_t equals some particular value (j), given the past, depends only on the most recent value of S_{t-1}. In other words,

$$P[S_t = j \mid S_{t-1} = i, S_{t-2} = m, ...] = P[S_t = j \mid S_{t-1} = i] = P_{ij},$$

where $\{P_{ij}\}_{i,j=1,2,...,K}$ are the one-step transition probabilities (Srinivasan and Mehata, 1978; Ross, 1996). The transition probability, P_{ij}, is the probability of transitioning from state i to state j in one time step. Note that $\sum_{j=1}^{K} P_{ij} = 1, P_{ij} \geq 0.$

Here, the output of the process is the set of states at each instant of time, where each state corresponds to an observable event. The above stochastic process is called an observable discrete time finite state Markov model.

2.2 Examples of hidden Markov models

In this section, we will give examples where the idea of the hidden Markov model (Rabiner, 1989; Elliott et al., 1995) is discussed and presented, in order to understand the concept of the HMM.

Examples:

1. A person is repeatedly rolling one of two dice picked at random, one of which is biased (unbalanced) and the other is unbiased (balanced). An observer records the results. If the dice are indistinguishable to the observer, then the two 'states' (i.e. biased dice or unbiased dice) in this model are hidden.
2. Consider an example of coin tossing. One person (person **A**) is in a room with a barrier (e.g., a curtain) through which he cannot see what is happening on the other side, where another person (person **B**) is performing a coin tossing experiment. Person **B** will tell person **A** the results of each coin flip. Person **A** only observes the results of the coin tosses, and he does not know anything about which coin gives the results. So the tossing experiment is hidden, providing a sequence of observations consisting of a series of heads and tails (T stands for tails and H stands for heads).

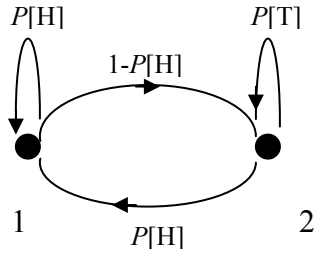
   For example : $Y_1\ Y_2 \ldots Y_T$

   H T … H.

Given the coin tossing experiment, the question of interest is how to build hidden Markov models that will explain the observation sequence. For example 2, we can consider several models: 1-coin model, 2-coins model and 3-coins model.

1-coin model:

Here, there are two states in the model, but each state is uniquely associated with either head (state 1) or tail (state 2); hence, this model is not hidden because the observation sequence uniquely defines the state.



Y = H H T T H T H H T T H……
S = 1 1 2 2 1 2 1 1 2 2 1…..

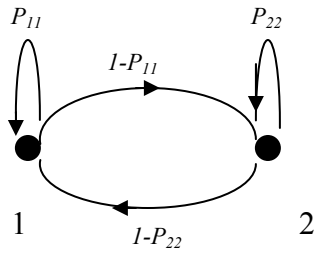
P[H]- The probability of observing a head
P[T]- The probability of observing a tail
1-*P*[H]- The probability of leaving state 1

Figure 2.1: 1- coin model

2-coins model:

There are two states in this model corresponding to a different, biased, coin being tossed; neither state is uniquely associated with either head or tail. Each state is characterized by a probability distribution of heads and tails, and the state transition matrix characterizes the transitions between the states. This matrix can be selected by a set of independent coin tosses or some other probabilistic event. The observable output sequences of 2-coins model are independent of the state transitions. This model is

hidden because we do not know exactly which coin (state) led to the head or tail of each observation.



P_{11}- The probability of staying in state 1

P_{22}- The probability of staying in state 2

$1\text{-}P_{11}$- The probability of leaving state 1

$1\text{-}P_{22}$- The probability of leaving state 2

Y = H H T T H T H H T T H……

S = 2 1 1 2 2 2 1 2 2 1 2…..

(1) P[H]=P_1 (2) P(H)=P_2

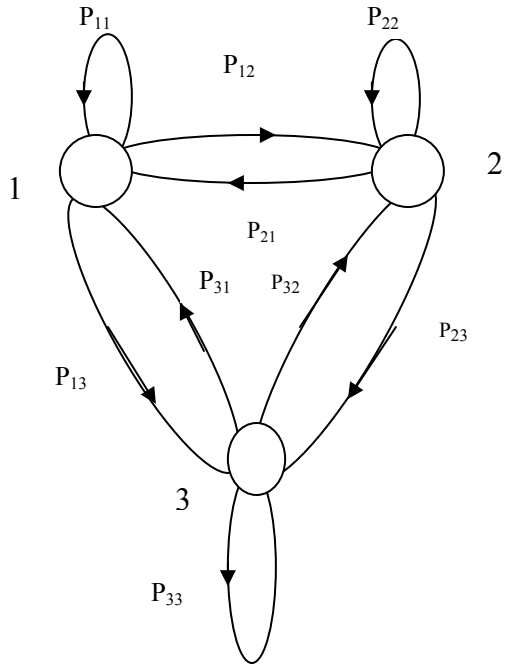
P[T]=$1\text{-}P_1$ P[T]=$1\text{-}P_2$,

where (1) is the probability distribution of heads and tails in state 1 and (2) is the probability distribution of heads and tails in state 2.

Figure 2.2: 2- coins model

3-coins model:

The third form of the HMM for explaining the observed sequence of coin tossing outcomes is given in Figure 2.3. This model corresponds to the example using three biased coins, and choosing from among the three based on some probabilistic event.



P_{11}- The probability of staying in state 1

P_{22}- The probability of staying in state 2

P_{33}- The probability of staying in state 3

P_{12}- The probability of leaving state 1 and reaching state 2

P_{21}- The probability of leaving state 2 and reaching state 1

P_{13}- The probability of leaving state1 and reaching state 3

P_{31}- The probability of leaving state 3 and reaching state 1

P_{32}- The probability of leaving state 3 and reaching state 2

P_{23}- The probability of leaving state 2 and reaching state 3

Y = H H T T H T H H T T H……

S = 3 1 2 3 3 1 1 2 3 1 3…..

(1) $P[H]=P_1$ (2) $P[H]=P_2$ (3) $P[H]=P_3$

$P[T]=1-P_1$ $P[T]=1-P_2$ $P[T]=1-P_3$,

where (1) is the probability distribution of heads and tails in state 1, (2) is the probability distribution of heads and tails in state 2 and (3) is the probability distribution of heads and tails in state 3.

Figure 2.3: 3- coins model

Sample calculation of hidden Markov model (HMM)

A hidden Markov model is defined by specifying following five things:

Q = the set of states = $\{1, 2, ..., k\}$.

V = the output observations = $\{v_1, v_2, ..., v_m\}$, where m is finite number.

$\pi(i)$ = Probability of being in state i at time $t = 0$ (i.e. in initial states).

A = transition probabilities = $\{P_{ij}\}$, where

P_{ij} =P[entering state j at time $t+1$ |in state i at time t] = $P[S_{t+1} = j \mid S_t = i]$.

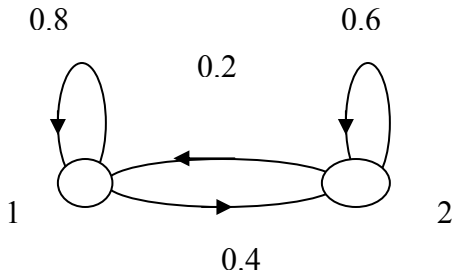
Note that the probability of going from state i to state j does not depend on the previous states at earlier times. This is called as Markov property.

B = output probabilities = $\{b_j(m)\}$,

where $\{b_j(m)\}$=P[producing v_m at time t | in state j at time t].

The above definition of a HMM applies to the special case where one has discrete states and discrete observations (Elliott et al., 1995).

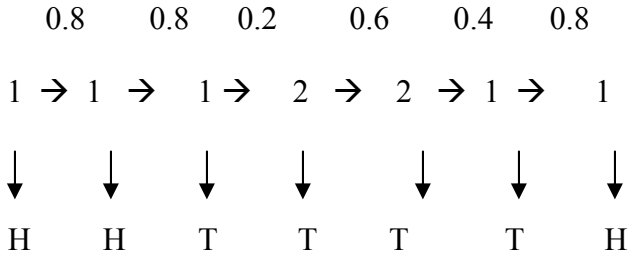
Consider the case with the 2- biased coins model in Figure 2.4. Here, two biased coins were flipped, and an observer was seeing the results of the coin flip but not which coin was flipped. The states of the HMM are 1 and 2 (two coins), the output observation is {H, T}, and transition and output probabilities are as labeled. Let the initial state probabilities are $\pi(1) = 1$ and $\pi(2) = 0$. This model has two states corresponding to two different coins. In state 1, the coin is biased strongly towards heads and in state 2; the coin is biased strongly towards tails. The state transition probabilities are 0.8, 0.6, 0.2, and 0.4.



(1) P [H]=2/3 (2) P [H]=1/6

P [T]=1/3 P [T]=5/6

Figure 2.4: 2-biased coins model



The probabilities for the following events can be calculated as follows:

1. The probability of the above state transition sequence:

   $P[1112211] = \pi(1)\, P_{11}P_{11}P_{12}P_{22}P_{21}P_{11} = 1\times 0.8\times 0.8\times 0.2\times 0.6\times 0.4\times 0.8 = 0.025.$

2. The probabilities of the above output sequence given the above transition sequence:

   $P[(\text{HHTTTTH})|(1112211)] = \frac{2}{3}\times\frac{2}{3}\times\frac{1}{3}\times\frac{5}{6}\times\frac{5}{6}\times\frac{1}{3}\times\frac{2}{3} = 0.023.$

3. The probability of the above output sequence and the above transition sequence:

   $P[(\text{HHTTTTH})\cap(1112211)] = 0.025\times 0.023 = 5.7\times 10^{-4}.$

In this case, the results of the coin flips and which coins are being flipped are known. In general, which coins are being flipped is unknown. That is, the underlying model is known and the output sequence is observed, while the state sequence is "hidden." In this case, the 2-biased-coins model is a hidden Markov model.

In the above examples, the outcomes of the tossed coins are T or H and only two observations are possible. More general situation is explained below: considering a set of N urns and each urn consisting of several colored balls (M).

The urn and ball model

Consider the situation where there are N urns in a room, and within each urn there are M distinct colours of balls (Figure 2.5). The physical process for obtaining observations is as follows. A person is in a room, and using some random process, he (or she) chooses an initial urn. From this urn, a ball is chosen at random, and its color is recorded as the observation. The ball is then replaced in the urn from which it was selected. A new urn is then selected according to the random selection process associated with the current urn, and the ball selection process is repeated. This entire process generates a finite observation sequence of colours, which can be considered as the observational output of the HMM. Here, each state corresponds to a specific urn and the ball colour probability is defined for each state. The choice of urns is dictated by the state transition matrix of the HMM.

Urn 1 Urn 2 … Urn N

P[Red]$=b_1(1)$ P[Red]$=b_2(1)$ P[Red]$=b_N(1)$

P[Blue]$=b_1(2)$ P[Blue]$=b_2(2)$ P[Blue]$=b_N(2)$

P[Green]$=b_1(3)$ P[Green]$=b_2(3)$ P[Green]$=b_N(3)$

…

P[Orange]$=b_1(M)$ P[Orange]$=b_2(M)$ P[Orange]$=b_N(M)$

The observation sequence is

Y= {Green, Green, Red, Yellow, Blue, …, Orange, Blue}

Figure 2.5: The Urn and Ball Model

2.3 Definition of the hidden Markov model

A hidden Markov Model is a doubly stochastic process, with an underlying stochastic process that is not observable (hidden), and can only be observed through another set of stochastic processes that produced the sequence of observations.

Simply stated, a hidden Markov model is a finite set of states, each of them being associated with a probability distribution, and the transition between the states being covered by the transition probability. In particular, the observation can be generated according to the associated probability distribution so it is only the outcome that is

observable, not the states; therefore, the states are "hidden" to the observer, hence the name "Hidden Markov Model" (Rabiner, 1989; Elliott et al., 1995).

To define the hidden Markov model completely we need to define the elements of the HMM:

1. The length of the observation sequence, T. So the states sequence can be written as $\{S_1, S_2, ..., S_T\}$ and the observation sequence would be $\{Y_1, Y_2, ..., Y_T\}$.
2. The number of states in the model, K. In the 2-coins model example, the states correspond to the choice of coins (i.e. two possible states). The state at time t is denoted as S_t throughout the thesis. In the Urn model, the number of states corresponds to the number of urns.
3. The number of distinct observation symbols per state, M. For the coin-tossing example, the observation symbols are simply the "H" and the "T". Considering the more general Urn model, the numbers of distinct observation symbols are M distinct colours.
4. A set of state transition probabilities $A = \{P_{ij}\}$,

   $P_{ij} = P[S_{t+1} = j \mid S_t = i], \quad 1 \le i, j \le K$,

   where S_t denotes the state at time t and P_{ij} denotes the transition probability which must satisfy the constraints

   $$P_{ij} \ge 0, \text{ for all } 1 \le i, j \le K$$

   $$\sum_{j=1}^{K} P_{ij} = 1, \text{ for all } 1 \le i \le K.$$

5. The probability distribution of the observation symbol in state j: $B = \{b_j(n)\}$

   $b_j(n) = P[v_n \text{ at time } t \mid S_t = j], \qquad 1 \le j \le K, \qquad 1 \le n \le M$,

   where v_n denotes the n^{th} observation symbol in a given state j.

   $b_j(n)$ should also satisfy the stochastic constraints

   $b_j(n) \ge 0$, $\quad 1 \le j \le K, \qquad 1 \le n \le M$ and

   $$\sum_{n=1}^{M} b_j(n) = 1, \ 1 \le j \le K .$$

6. The above probability distribution is the case when the observations are discrete. The initial state distribution $\pi = \{\pi_i\}$, where

   $$\pi_i = P[S_1 = i] \ , \ 1 \le j \le K .$$

From above definitions, it is clear that a complete specification of an HMM involves three model parameters (K, M, T) and three sets of probability parameters (A, B, π). Therefore, for convenience, we can use the compact notation $\lambda = (A, B, \pi)$ to denote the complete set of parameters of the model throughout the thesis.

Before we go further, there are some assumptions that are made in the theory of hidden Markov models for mathematical and computational tractability. First, it is assumed that the next state is dependent only on the current state, which is called the Markov assumption. That is,

$$P[S_{t+1} = j \mid S_t = i_t, S_{t-1} = i_{t-1}, ..., S_0 = i_0] = P[S_{t+1} = j \mid S_t = i_t] .$$

Second, there is the homogeneity assumption (i.e. state transition probabilities are independent of the actual time at which the transition takes place)

$$P[S_{t_1+1} = j \mid S_{t_1} = i] = P[S_{t_2+1} = j \mid S_{t_2} = i].$$

Third, the statistical independence of observations, i.e. suppose we have a sequence of observations $\mathbf{Y} = \{Y_1, Y_2, ..., Y_T\}$, and the sequence of states $\{S_1, S_2, ..., S_T\}$ then the probability distribution of generating the current observation depends only on the current state. That is,

$$P[\mathbf{Y} = \mathbf{y} \mid S_1 = i_1, ..., S_T = i_T; \lambda] = \prod_{t=1}^{T} P[Y_t = y_t \mid S_t = i_t; \lambda],$$

and these assumptions are used to solve the problems associated with hidden Markov models.

2.4 Definition of the hidden Markov random field model

In this section, the Markov random field model is introduced, followed by the definition of the hidden Markov random model (Kunsch, 1995; Elliott et al., 1995 and 1996; Fishman, 1996).

2.4.1 Markov random fields

A random field is a stochastic process defined on a two-dimensional set, that is, a region of the plane, or a set of even higher dimension. The two-dimensional case will be the

focus of this section. Random fields, which possess a Markov property, are called Markov random fields (Elliott et al., 1995 and 1996).

Let us generalize this idea in a two-dimensional setting. Let $\mathbf{Z}$ be the set of integers, and let $\mathbf{S} \subset \mathbf{Z}^2$ be a finite rectangular two-dimensional lattice of integer points. Typically, it will take $\mathbf{S} = \{0,1,...,n-1\} \times \{0,1,...,m-1\}$, for some n and m. $\mathbf{S}$ is a two-dimensional lattice containing $n \times m$ points. The points in $\mathbf{S}$ are often called sites. To define a Markov structure on the set $\mathbf{S}$, we define what is meant by two points being neighbours. Different definitions may suit different purposes, or applications. However, the following two general conditions should include in the definition.

(i) A site must be a neighbour of itself.

(ii) If t is neighbour of s, then s is a neighbour of t.

The second condition is a symmetry requirement. It can be written $s \sim t$ if the sites $s, t \in \mathbf{S}$ are neighbours. Two common neighbourhood structures are given in Figure 2.6. If s is a site, the neighbourhood $\mathbf{N}_s$ of s can be defined as the set of all its neighbours; $\mathbf{N}_s = \{t \in \mathbf{S} : t \sim s\}$. Hence, Figure 2.6 illustrates the neighbourhood of the middle site, for two different structures. In these structures, special care must be taken at the edge of the lattice $\mathbf{S}$, since sites located there have smaller neighbourhoods. One way of defining the neighbourhood structure is "wrapping around" the lattice and define sites at the other end of the lattice as neighbours.

Concept of clique:

Cliques are particular subsets of the sites in $\mathbf{S}$, defined in the following way:

(i) Any single site *s* is a clique.

(ii) Any subset $\mathbf{C} \subset \mathbf{S}$ of more than one site is a clique if all pairs of sites in $\mathbf{C}$ are neighbours.

Hence, what the cliques look like depends on the neighbourhood system. Figure 2.6 shows what cliques there are for the two neighbourhood systems displayed therein. Note that these schematic cliques should be moved around over the lattice to find out all the subsets of sites that fit with the given pattern.

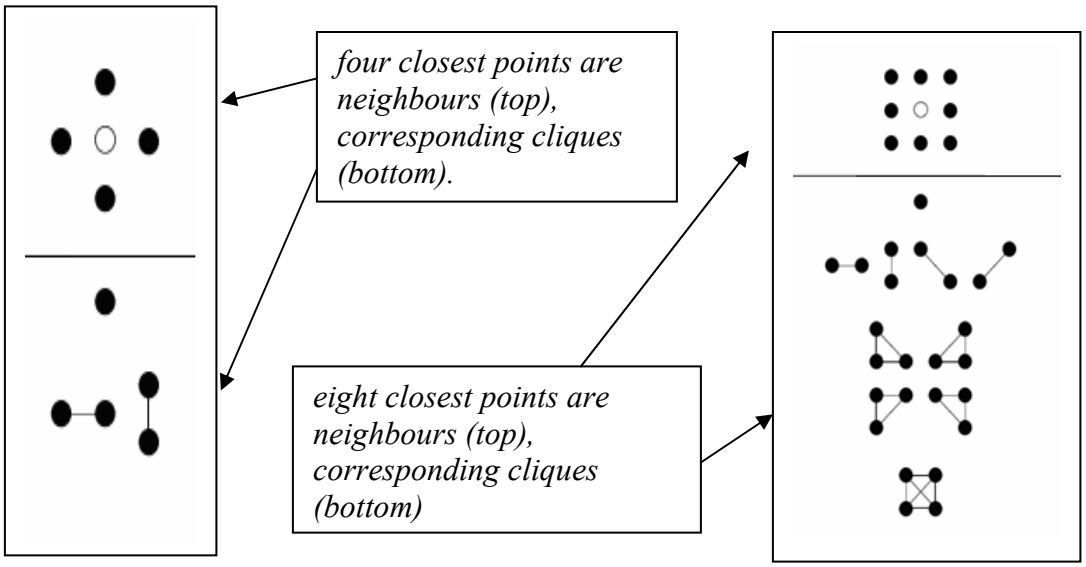


Figure 2.6: Two different neighbourhood structures and their corresponding cliques

Now consider a random field $\{X(\mathbf{s}):\mathbf{s}\in\mathbf{S}\}$ defined on $\mathbf{S}$, that is, a collection $X(\mathbf{s})$ of random variables indexed by sites in $\mathbf{S}$. These random variables are assumed to take their values in a finite set χ, the state space. Some examples of χ are $\chi=\{-1,+1\}$ and $\chi=\{1,2,...,r\}$. The set $\chi^{\mathbf{S}}$ is the set of elements of the form $x=\{x(\mathbf{s}):\mathbf{s}\in\mathbf{S}\}$ with $x(\mathbf{s})\in\chi$ for each *s*. An element of $\chi^{\mathbf{S}}$ will often called as a configuration (of the random field). Also often we can simply write this as $\mathbf{X}$ for $\{X(\mathbf{s}):\mathbf{s}\in\mathbf{S}\}$ and think of $\mathbf{X}$ as a random variable with values in $\chi^{\mathbf{S}}$, the set of configurations. Letting $|\mathbf{S}|$ denote the number of elements of $\boldsymbol{S}$ and similarly for χ, the number of elements of the configuration space $\chi^{\mathbf{S}}$ is $|\chi|^{|\mathbf{S}|}$ and it is hence often extremely large. For example, if $\chi=\{-1,+1\}$ and $\mathbf{S}$ is a lattice of size 128×128, its size is 2^{128^2}. If $\mathbf{A}$ is a subset of $\mathbf{S}$, write $X(\mathbf{A})$ for $\{X(\mathbf{s}):\mathbf{s}\in\mathbf{A}\}$, that is the collection of random variables on $\mathbf{A}$, and similarly for a particular configuration $x=\{x(\mathbf{s}):\mathbf{s}\in\mathbf{S}\}$. The symbol \ denotes set-difference; for example, $\mathbf{S}\setminus\{\mathbf{s}\}$ is the set of sites in $\mathbf{S}$ except $\mathbf{s}$, and write this difference as $\mathbf{S}\setminus\mathbf{s}$. Now the random field $\{X(\mathbf{s}):\mathbf{s}\in\mathbf{S}\}$ is a Markov random field (MRF) on $\mathbf{S}$ (with respect to the given neighbourhood structure) if

$P[X(\mathbf{s})=x(\mathbf{s})\,|\,X(\mathbf{S}\setminus\mathbf{s})=x(\mathbf{S}\setminus\mathbf{s})]=P[X(\mathbf{s})=x(\mathbf{s})\,|\,X(\mathbf{N_s})=x(\mathbf{N_s})]$ for all sites $\mathbf{s}\in\mathbf{S}$ and all configurations $x\in\chi^{\mathbf{S}}$. In other words, the distribution of $X(\mathbf{s})$, given all other sites, depends at the realized values in its neighbourhood only. These conditional distributions are often called the local specification of the MRF. Two examples are presented to get an idea of different MRF:

The Ising model

Assuming $\chi = \{-1,+1\}$ and the neighbourhood structure to the left in Figure 2.6, an often used local specification is

$$P[X(\mathbf{s}) = x(\mathbf{s}) \mid X(\mathbf{N}_\mathbf{s}) = x(\mathbf{N}_\mathbf{s})] = \frac{\exp(\beta \sum_{\mathbf{t} \in \mathbf{N}_\mathbf{s}} x(\mathbf{s})x(\mathbf{t}))}{\exp(-\beta \sum_{\mathbf{t} \in \mathbf{N}_\mathbf{s}} x(\mathbf{t})) + \exp(\beta \sum_{\mathbf{t} \in \mathbf{N}_\mathbf{s}} x(\mathbf{t}))}$$

for some real β; note that the denominator does not depend on $x(\mathbf{s})$ and is only a normalizing factor to make the right hand side a proper distribution, summing to unity. This model is called the Ising model (McCoy et al., 1973; Binder, 1979; Binder et al., 1992), after German physicist Ising who invented it with the original purpose of using it as an idealized model of a ferromagnetic material. The sum in the exponent is positive if $x(\mathbf{s})$ has the same sign as the most of its neighbours. Hence, if β>0 the sites interact such that configurations x with many neighbours of the same sign will have large probabilities. On the contrary, if β<0, configurations with many neighbours having opposite signs will have large probabilities.

The Potts model

If there is no particular assumption on χ, except it being finite, and any of the neighbourhood systems of Figure 2.6, or some other one, a possible local specification is

$$P[X(\mathbf{s}) = x(\mathbf{s}) \mid X(\mathbf{N}_\mathbf{s}) = x(\mathbf{N}_\mathbf{s})] = \frac{\exp(\beta \#\{\mathbf{t} \in \mathbf{N}_\mathbf{s} : x(\mathbf{t}) \neq x(\mathbf{s})\})}{\sum_{i \in \chi} \exp(\beta \#\{\mathbf{t} \in \mathbf{N}_\mathbf{s} : x(\mathbf{t}) \neq i\})}$$

for some real β. Again, the denominator does not depend on $x(\mathbf{s})$ and is only a normalizing factor. This model is called the Potts model (Binder, 1979; Wu, 1982; Binder et al., 1992). Note that, $\#\{\mathbf{t} \in \mathbf{N}_\mathbf{s} : x(\mathbf{t}) \neq x(\mathbf{s})\}$ is the number of neighbours of $\mathbf{s}$ that have values different from $x(\mathbf{s})$. Hence, if β>0 this model gives large probabilities to configurations x in which there are many neighbours with different values. If β<0, the model works the opposite way, that is, configurations with many neighbours with equal values have large probabilities.

So far the local specification of a MRF is discussed, and it is also interesting to find out a corresponding distribution on $\chi^\mathbf{S}$, that is, in the probabilities of various configurations x. This distribution can be denoted by π; hence,

$$\pi(x) = P[X = x] = P[X(\mathbf{s}) = x(\mathbf{s}), \forall \mathbf{s} \in \mathbf{S}]$$

for any configuration $x \in \chi^\mathbf{S}$. Now assume for each clique C there is a function $V_\mathbf{C} : \chi^\mathbf{S} \to \mathbf{R}$. That is, $V_\mathbf{C}$ maps a configuration x into a real number. Moreover, $V_\mathbf{C}$ must not depend on sites other than those in $\mathbf{C}$. This can be written as $V_\mathbf{C}(x) = V_\mathbf{C}(X(\mathbf{C}))$. A probability mass function, or distribution, π on the configuration space $\chi^\mathbf{S}$ of the form

$$\pi(x) = Z^{-1} \exp\left\{\sum_\mathbf{C} V_\mathbf{C}(x)\right\}$$

is called a Gibbs distribution. Here the sum runs over all cliques C. The energy function is defined as $U(x) = \sum_\mathbf{C} V_\mathbf{C}(x)$ which is a sum of clique potentials $V_\mathbf{C}(x)$ over all possible cliques $\mathbf{C}$ and the normalizing constant (or partition function) Z is given by

$$Z = \sum_{x \in \chi^S} \exp\left\{ \sum_{C} V_C \right\}$$

and is generally infeasible to compute as the outer sum runs over a very large set. The importance of Gibbs distributions is made clear from the following facts:

(i) Any random field with a distribution π which is a Gibbs distribution is a Markov random field with respect to the neighbourhood system governing the cliques.

(ii) Any random field which is Markov with respect to a give neighbourhood system has a distribution π, which is a Gibbs distribution generated by the corresponding cliques.

Hence, according to the Hammersley-Clifford theorem (Fishman, 1996), an MRF can equivalently be characterized by a Gibbs distribution. For more details on the MRF and the Gibbs distribution, see Geman and Geman (1984).

2.4.2 Hidden Markov random field (HMRF) model

The concept of a hidden Markov random field model (Elliott et al., 1995 and 1996) is derived from the hidden Markov model, which is defined as stochastic processes generated by a Markov chain whose state sequence cannot be observed directly, only through a sequence of observations. Each observation is assumed to be a stochastic function of the state sequence. The underlying Markov chain changes its state according to a $\ell \times \ell$ transition probability matrix, where ℓ is the number of states.

Since original HMMs were designed as one-dimensional Markov chains with first-order neighborhood systems, it cannot directly be used in two-dimensional problems such as image segmentation. A special case of an HMM, in which the underlying stochastic process is an MRF instead of a Markov chain, is referred to as a hidden Markov random field model (Zhang et al., 2001). Mathematically, an HMRF model is characterized by the following:

- Hidden Markov Random Field (HMRF)

  The random field $\mathbf{X} = \{X(\mathbf{s}) : \mathbf{s} \in \mathbf{S}\}$ is an underlying HMRF assuming values in a finite state space $L = (1,....,\ell)$ with probability distribution π. The state of $\mathbf{X}$ is unobservable.

- Observable Random Field

  $\mathbf{Y} = \{Y(\mathbf{s}) : \mathbf{s} \in \mathbf{S}\}$ is a random field with a finite state space $D = (1,...,d)$. Given any particular configuration $\mathbf{x} \in \boldsymbol{\chi}$, every $Y(\mathbf{s})$ follows a known conditional probability distribution $p(y(\mathbf{s}) \mid x(\mathbf{s}))$ of the same functional form $f(y(\mathbf{s}); \boldsymbol{\theta}_{x(\mathbf{s})})$, where $\boldsymbol{\theta}_{x(\mathbf{s})}$ are the involved parameters. This distribution is called the emission probability function and $\mathbf{Y}$ is also referred to as the emitted random field.

- Conditional Independence

  For any $\mathbf{x} \in \boldsymbol{\chi}$, the random variables $Y(\mathbf{s})$ are conditional independent

$$p(\mathbf{y} \mid \mathbf{x}) = \sum_{\mathbf{s} \in \mathbf{S}} p(y(\mathbf{s}) \mid x(\mathbf{s})).$$

Based on the above, the joint probability of $(\mathbf{X}, \mathbf{Y})$ can be written as

$$p(\mathbf{y}, \mathbf{x}) = p(\mathbf{y} \mid \mathbf{x}) p(\mathbf{x}) = p(\mathbf{x}) \sum_{\mathbf{s} \in \mathbf{S}} p(y(\mathbf{s}) \mid x(\mathbf{s})).$$

According to the local characteristics of MRFs, the joint probability of any pair of $(X(\mathbf{s}), Y(\mathbf{s}))$, given $X(\mathbf{s})$'s neighborhood configuration $X(\mathbf{N}_\mathbf{s})$, is

$$p(y(\mathbf{s}), x(\mathbf{s}) \mid x(\mathbf{N}_\mathbf{s})) = p(y(\mathbf{s}) \mid x(\mathbf{s}))\, p(x(\mathbf{s}) \mid x(\mathbf{N}_\mathbf{s})).$$

The marginal probability distribution of $Y(\mathbf{s})$ dependent on the parameter set $\boldsymbol{\theta}$ and $X(\mathbf{N}_\mathbf{s})$ can be written as

$$p(y(\mathbf{s}) \mid x(\mathbf{N}_\mathbf{s}), \boldsymbol{\theta}) = \sum_{\ell \in L} p(y(\mathbf{s}), l \mid x(\mathbf{N}_\mathbf{s}), \boldsymbol{\theta})$$

$$= \sum_{\ell \in L} f(y(\mathbf{s}); \theta_\ell)\, p(\ell \mid x(\mathbf{N}_\mathbf{s})) \text{ where } \boldsymbol{\theta} = \{\theta_\ell : \ell \in L\}.$$

This model is called the hidden Markov random field model. Note that the concept of an HMRF is different from that of an MRF in the sense that the former is defined with respect to a pair of random variable families, $(\mathbf{X}, \mathbf{Y})$ while the latter is only defined with respect to $\mathbf{X}$.

CHAPTER 3

INFERENCE IN HIDDEN MARKOV MODELS

3.1 Introduction

Given the HMM model in Chapter 2, there are three basic computational problems that are useful for solving real world problems. The three problems are as follows:

Problem 1: Given the observation sequence $\mathbf{Y} = \{Y_1, Y_2, ..., Y_T\}$, and the model $\lambda = (A, B, \pi)$, how do we compute $P[\mathbf{Y} = \mathbf{y}; \lambda]$, the probability or likelihood of occurrence of the observation sequence $\mathbf{Y} = \{Y_1, Y_2, ..., Y_T\}$ given the parameter set λ ?

We can consider problem 1 as an evaluation problem, namely given a model and a sequence of observations, how do we compute the probability that the model produced the observed sequence. We can also view this problem as how well the given model matches a given observation sequence. For example, if we are trying to choose among several computing models, the solution to problem 1 allows us to choose the model which best matches the observations (Rabiner, 1989).

Problem 2: Given the observation sequence $\mathbf{Y} = \{Y_1, Y_2, ..., Y_T\}$, and the model $\lambda = (A, B, \pi)$, how do we choose a state sequence $\mathbf{S} = \{S_1, S_2, ..., S_T\}$ so that $P[\mathbf{Y} = \mathbf{y}, \mathbf{S} = \mathbf{s}; \lambda]$, the joint probability of the observation sequence $\mathbf{Y} = \{Y_1, Y_2, ..., Y_T\}$ and the state sequence given the model is maximized.

Problem 2 is the one in which we attempt to discover the hidden part of the model, that is, to find the "correct" state sequence. In practical situations, we usually use an optimality criterion to solve this problem as best as possible, since there is no "correct" state sequence to be found.

Problem 3: How do we estimate the hidden Markov model parameters $\lambda = (A, B, \pi)$ so that $P[\mathbf{Y} = \mathbf{y}; \lambda]$ (or $P[\mathbf{Y} = \mathbf{y}, \mathbf{S} = \mathbf{s}; \lambda]$) is maximized given the model?

Problem 3 is to determine a method to adjust the models parameters to maximize the probability of the observation sequence given the model. The maximization of the probability function can be done using an iterative procedure or using gradient techniques.

3.2 Solutions to three estimation problems:

3.2.1 Problem 1 and its solution

Problem 1 is the evaluation problem; that is, given the model and a sequence of observations, how we can compute the probability that the model produced the observed

sequence. If we have several competing models, a solution to problem 1 allows us to choose the model which best matches the observations.

A most straightforward way to determine $P[\mathbf{Y}=\mathbf{y};\lambda]$ is to find out $P[\mathbf{Y}=\mathbf{y},\mathbf{S}=\mathbf{s};\lambda]$ for a fixed state sequence $\mathbf{S}=\{S_1,S_2,...,S_T\}$ then multiply it by $P[\mathbf{S}=\mathbf{s};\lambda]$ and then sum up over all possible states $\mathbf{S}$.

We have a model λ and a sequence of observations $\mathbf{Y}=\{Y_1,Y_2,...,Y_T\}$ where T is the number of observations and we want to find the probability of the observation sequence $P[\mathbf{Y}=\mathbf{y};\lambda]$ given the model. One could calculate $P[\mathbf{Y}=\mathbf{y};\lambda]$ through enumerating every possible state sequence of length T. Hence

$$P[\mathbf{Y}=\mathbf{y};\lambda]=\sum_{\forall S}P[\mathbf{Y}=\mathbf{y}\mid\mathbf{S}=\mathbf{s};\lambda]P[\mathbf{S}=\mathbf{s};\lambda],\quad \text{where } \mathbf{S}=\{S_1,S_2,...,S_T\}$$

$$=\sum_{S_1,S_2,...,S_T}\pi_{S_1}b_{S_1}(y_1)P_{S_1S_2}b_{S_2}(Y_2)...P_{S_{T-1}S_T}b_{S_T}(y_T). \qquad (3.1)$$

But this calculation for $P(\mathbf{Y}=\mathbf{y};\lambda)$ according to (3.1), involves several operations of the order of $2TK^T$, which is very large even if the length of the sequence, T, is moderate. So another procedure must be applied to solve problem 1. Fortunately, this procedure, the forward procedure, exists and calculates this quantity in a moderate time (Baum et al., 1967; Rabiner, 1989).

The forward variable $\alpha_t(j)$ is defined as the probability of $\mathbf{Y}^{(t)}$, the partial observation sequence $\mathbf{Y}^{(t)}=\{Y_1, Y_2,..., Y_t\}$, when it terminates at state j given the hidden Markov model parameters λ. Thus,

$$\alpha_t(j) = P[\mathbf{Y}^{(t)} = \mathbf{y}^{(t)}, S_t = j; \lambda], \qquad j = 1, 2,..., K. \tag{3.2}$$

then $$P[\mathbf{Y} = \mathbf{y}; \lambda] = \sum_{j=1}^{K} P[\mathbf{Y}^{(t)} = \mathbf{y}^{(t)}, S_t = j; \lambda], \quad 1 \le t \le T$$

$$= \sum_{j=1}^{K} \alpha_t(j).$$

One can solve for $\alpha_t(j)$ inductively, through the equation:

$$\alpha_t(j) = P[\mathbf{Y}^{(t)} = \mathbf{y}^{(t)}, S_t = j]$$

$$= \sum_{i=1}^{K} P[Y_t = y_t, \mathbf{Y}^{(t-1)} = \mathbf{y}^{(t-1)}, S_t = j, S_{j-1} = i].$$

Using the Bayes law and the independence assumption one can obtain the following:

$$= \sum_{i=1}^{K} P[\mathbf{Y}^{(t-1)} = \mathbf{y}^{(t-1)}, S_{t-1} = i] P[Y_t = y_t, S_t = j \mid \mathbf{Y}^{(t-1)} = \mathbf{y}^{(t-1)}, S_{t-1} = i]$$

$$= \sum_{i=1}^{K} P[\mathbf{Y}^{(t-1)} = \mathbf{y}^{(t-1)}, S_{t-1} = i] P[S_t = j \mid \mathbf{Y}^{(t-1)} = \mathbf{y}^{(t-1)}, S_{t-1} = i] P[Y_t = y_t \mid S_t = j, \mathbf{Y}^{(t-1)} = \mathbf{y}^{(t-1)}, S_{t-1} = i]$$

$$= \sum_{i=1}^{K} P[\mathbf{Y}^{(t-1)} = \mathbf{y}^{(t-1)}, S_{t-1} = i] P[S_t = j \mid S_{t-1} = i] P[Y_t = y_t \mid S_t = j]$$

$$= \sum_{i=1}^{K} [\alpha_{t-1}(i) P_{ij}] b_j(y_t).$$

Therefore,

$$\alpha_t(j) = b_j(y_t) \sum_{i=1}^{K} \alpha_{t-1}(i) P_{ij}, \qquad 1 \le t \le T, \qquad 1 \le j \le K, \tag{3.3}$$

with

$$\alpha_1(j) = P[Y_1 = y_1, S_t = j] = \pi_j b_j(y_1).$$

Using this equation we can calculate $\alpha_T(j)$, $1 \le j \le K$, and then

$$P[\mathbf{Y} = \mathbf{y}; \lambda] = \sum_{j=1}^{K} \alpha_T(j). \tag{3.4}$$

This method is called the forward method and requires a calculation of the order K^2T, rather than $2TK^T$, as required by the direct calculation previously mentioned.

As an alternative to the forward procedure, there exists a backward procedure (Baum et al., 1967; Rabiner, 1989), which is able to solve $P[\mathbf{Y} = \mathbf{y}; \lambda]$. In a similar way, the backward variable $\beta_t(i)$ can be defined as

$$\beta_t(i) = P[\mathbf{Y}^{*(t)} = \mathbf{y}^{*(t)} \mid S_t = i; \lambda], \tag{3.5}$$

where $\mathbf{Y}^{*(t)}$ denotes $\{Y_{t+1}, Y_{t+2}, ..., Y_T\}$ (i.e. the probability of the partial observation sequence from *t+1* to *T* given the current state *i* and the model λ).

Note that

$$\beta_{T-1}(i) = P[\mathbf{Y}^{*(T-1)} = \mathbf{y}^{*(T-1)} \mid S_{T-1} = i; \lambda]$$

$$= P[Y_T = y_T; S_{T-1} = i] = \sum_{i=1}^{K} P_{ij} b_j(y_T). \tag{3.6}$$

As for of $\alpha_t(j)$, one can solve for $\beta_t(i)$ inductively and can get the following recursive relationship.

Now,

first initialize $\beta_T(i) = 1, 1 \le i \le K.$ (3.7)

Then for $t = T-1, T-2, ..., 2, 1$ and $1 \le i \le K$,

$$\beta_{t-1}(i) = P[\mathbf{Y}^{*(t-1)} = \mathbf{y}^{*(t-1)} \mid S_{t-1} = i]$$

$$= P[Y_t = y_t, \mathbf{Y}^{*(t)} = \mathbf{y}^{*(t)} \mid S_{t-1} = i]$$

$$= \sum_{j=1}^{K} P[Y_t = y_t, \mathbf{Y}^{*(t)} = \mathbf{y}^{*(t)}, S_t = j \mid S_{t-1} = i]$$

$$= \sum_{j=1}^{K} P[\mathbf{Y}^{*(t)} = \mathbf{y}^{*(t)} \mid S_t = j] P[Y_t = y_t \mid S_t = j] P[S_t = j \mid S_{t-1} = i]$$

$$= \sum_{j=1}^{K} \beta_t(j) b_j(y_t) P_{ij}$$

$$\beta_{t-1}(i) = \sum_{j=1}^{K} P_{ij} b_j(y_t) \beta_t(j), \quad 1 \le i \le K, \ \ 1 \le t \le T-1. \tag{3.8}$$

Finally it can be demonstrated that

$$P[\mathbf{Y} = \mathbf{y}; \lambda] = P[\mathbf{Y}^{*(0)} = \mathbf{y}^{*(0)}; \lambda]$$

$$= P[Y_1 = y_1, Y_2 = y_2, ..., Y_T = y_T]$$

$$= \sum_{i=1}^{K} P[Y_1 = y_1, \mathbf{Y}^{*(1)} = \mathbf{y}^{*(1)} \mid S_1 = i]$$

$$= \sum_{i=1}^{K} P[\mathbf{Y}^{*(1)} = \mathbf{y}^{*(1)} \mid S_1 = i] P[Y_1 = y_1 \mid S_1 = i]$$

$$= \sum_{i=1}^{K} P[\mathbf{Y}^{*(1)} = \mathbf{y}^{*(1)} \mid S_1 = i] \pi_i b_1(y_1)$$

$$P[\mathbf{Y} = \mathbf{y}; \lambda] = \sum_{i=1}^{K} \beta_1(i) \pi_i b_1(y_1) . \tag{3.9}$$

3.2.2 Problem 2 and its solution

Given a sequence of observations $\mathbf{Y} = \{Y_1, Y_2, ..., Y_T\}$ and the model λ, we want to find the most likely state sequence associated with the given observation sequence.

The solution to this problem depends on the way "the most likely state sequence" is defined. One method is to find the most likely state S_t at time t and to concatenate all such S_t's. However, sometimes this approach does not give a physically meaningful state sequence. The most widely used criterion is to maximize $P[\mathbf{Y} = \mathbf{y}, \mathbf{S} = \mathbf{s}; \lambda]$. That is, to maximize the probability of observing observation sequence $\mathbf{Y} = \{Y_1, Y_2, ..., Y_T\}$ and the state sequence $\mathbf{S} = \{S_1, ..., S_T\}$ given their joint distribution $f(\mathbf{y}, \mathbf{s})$.

Since the model $\lambda = (A, B, \pi)$ and the observation sequence is $\mathbf{Y} = \{Y_1, Y_2, ..., Y_T\}$, the probability of the state path and observation sequence given the model would be:

$$P[\mathbf{Y} = \mathbf{y}, \mathbf{S} = \mathbf{s}; \lambda] = P[\mathbf{Y} = \mathbf{y} \mid \mathbf{S} = \mathbf{s}; \lambda] P[\mathbf{S} = \mathbf{s}; \lambda]$$

$$= \pi_{S_1} b_{S_1}(y_1) P_{S_1 S_2} b_{S_2}(y_2) ... P_{S_{T-1} S_T} b_{S_T}(y_T). \tag{3.10}$$

To write this in a form of summations, we define $U(\mathbf{s})$ as

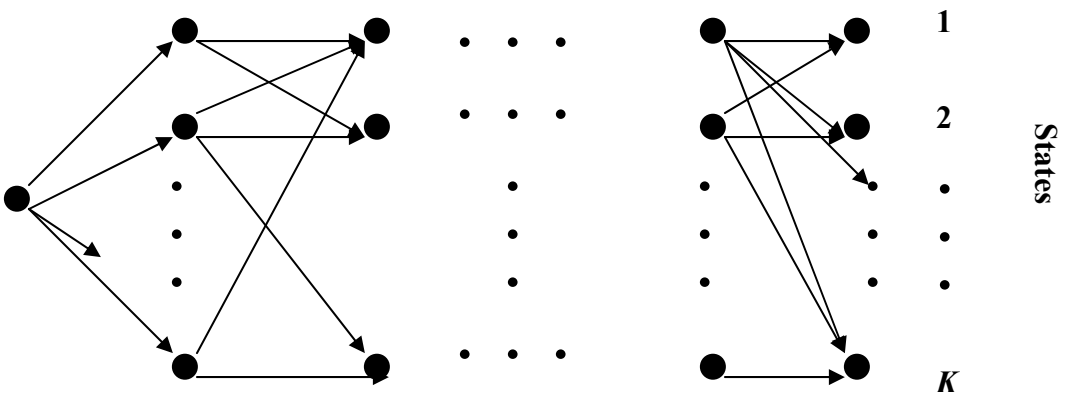
$$U(\mathbf{s}) = -\ln(P[\mathbf{Y} = \mathbf{y}, \mathbf{S} = \mathbf{s}; \lambda])$$

$$= -[\ln(\pi_{S_1} b_{S_1}(y_1)) + \sum_{t=2}^{T} \ln(P_{S_{t-1} S_t} b_{S_t}(y_t))]. \tag{3.11}$$

Since ln() is monotonic function if $-\ln(P[\mathbf{Y} = \mathbf{y}, \mathbf{S} = \mathbf{s}; \lambda])$ is minimum, then this will give us the state sequence for $P[\mathbf{Y} = \mathbf{y}, \mathbf{S} = \mathbf{s}; \lambda]$ is maximum. Therefore,

$$P[\mathbf{Y}=\mathbf{y},\mathbf{S}=\mathbf{s}_{opt};\lambda]=\max_{\mathbf{s}} P[\mathbf{Y}=\mathbf{y},\mathbf{S}=\mathbf{s};\lambda] \Leftrightarrow U(\mathbf{s}_{opt})=\min_{\mathbf{s}} U(\mathbf{s})\cdot$$

By starting at the unique point in time 0 and moving from a point in time t to a point in time $t+1$ in an optimal way, the distance between points in time t and points in time $t+1$ are equal to $-\ln(P_{S_{t-1}S_t} b_{S_t}(y_t))$ for $t \geq 1$. This distance is associated to the transition from state S_{t-1} to S_t. The problem of finding the most likely state sequence associated with the given observation sequence is then the shortest path in a following grid of points.



In this graph, the vertex corresponds to the states and the length between two vertexes is proportional to the weight on the edge (not shown in the graph). Finding the shortest-path problem is one of the most fundamental problems in graph theory and can be solved by dynamic programming approaches, such as the Viterbi Algorithm (Forney, 1973). With the research paper written by A.J. Viterbi in 1967, the Viterbi Algorithm made its first appearance in the coding literature.

Letting $U_t(S_1 = i_1, S_2 = i_2, ..., S_t = i_t)$ be the first t terms of $U(\mathbf{s})$ and $V_t(i_t)$ be the minimal accumulated distance when we are in state i at time t,

$$U_t(S_1 = i_1, S_2 = i_2, \ldots, S_t = i_t) = -[\ln(\pi_{S_1} b_{S_1}(y_1)) + \sum_{i=2}^{t} \ln(P_{S_{i-1}Si} b_{S_i}(y_i))]$$

$$V_t(i_t) = \min_{S_1, S_2, \ldots S_{t-1}, S_t = i_t} U_t(S_1 = i_1, S_2 = i_2, \ldots, S_{t-1} = i_{t-1}, S_t = i_t).$$

Viterbi algorithm can be carried out by following four steps:

1. Initialize the $V_1(i_1)$ for all $1 \le i \le K$:

   $$V_1(i_1) = -\ln(\pi_{S_i} b_{S_i}(y_{i_1})).$$

2. Inductively calculate the $V_t(i_t)$ for all $1 \le i_t \le K$, from time $t = 2$ to $t = T$:

   $$V_t(i_t) = \min_{1 \le i_{t-1} \le K} [V_{t-1}(i_{t-1}) - \ln(P_{S_j S_i} b_{S_i}(y_{i_t})].$$

3. Then we get the minimal value of $U(\mathbf{s})$:

   $$\min_{i_1, i_2, \ldots, i_T} U(\mathbf{s}) = \min_{1 \le i_T \le K} [V_T(i_T)].$$

4. Finally we trace back the calculation to find the optimal state path $\mathbf{S}_{opt} = \{S_{1,opt}, S_{2,opt}, \ldots, S_{T,opt}\}$.

3.2.3 Problem 3 and its solution

This problem is concerned with how to determine a way to adjust the model parameters so that the probability of the observation sequence given the model is maximized. However, there is no known way to solve for the model analytically and maximize the probability of the observation sequence. The iterative procedures such as the Baum-Welch Method (equivalently the EM (expectation-maximization) method (Dempster et

al., 1977) or gradient techniques can be used to estimate the model parameters. We will describe the solution for this problem based on the Baum-Welch method.

Baum-Welch algorithm

To describe the Baum-Welch algorithm (forward-backward algorithm) one needs to define two other variables in addition to the forward and backward variables defined previously.

The first variable is defined as the probability of being in state i at time t, and in state j at time $t+1$, given the model and the observation sequence

$$\xi_t(i,j) = P[S_t = i, S_{t+1} = j \mid \mathbf{Y} = \mathbf{y}; \lambda]. \tag{3.12}$$

Using Bayes law and the independency assumption, the equation (3.12) can be written as

$$\begin{aligned}
\xi_t(i,j) &= \frac{P[S_t = i, S_{t+1} = j, \mathbf{Y} = \mathbf{y}; \lambda]}{P[\mathbf{Y} = \mathbf{y}; \lambda]} \\
&= \frac{P[S_t = i, \mathbf{Y}^{(t)} = \mathbf{y}^{(t)}; \lambda] P[\mathbf{Y}^{*(t)} = \mathbf{y}^{*(t)}, S_{t+1} = j \mid S_t = i; \lambda]}{P[\mathbf{Y} = \mathbf{y}; \lambda]} \\
&= \frac{P[S_t = i, \mathbf{Y}^{(t)} = \mathbf{y}^{(t)}; \lambda] P[S_{t+1} = j \mid S_t = i] P[\mathbf{Y}^{*(t)} = \mathbf{y}^{*(t)} \mid S_{t+1} = j, S_t = i; \lambda]}{P[\mathbf{Y} = \mathbf{y}; \lambda]} \\
&= \frac{P[S_t = i, \mathbf{Y}^{(t)} = \mathbf{y}^{(t)}; \lambda] P[S_{t+1} = j \mid S_t = i] P[Y_{t+1} = y_{t+1} \mid S_{t+1} = j; \lambda] P[\mathbf{Y}^{*(t+1)} = \mathbf{y}^{*(t+1)} \mid S_{t+1} = j; \lambda]}{P[\mathbf{Y} = \mathbf{y}; \lambda]}
\end{aligned} \tag{3.13}$$

and by the way that forward and backward variables are defined, we can use them to write $\xi_t(i,j)$ in the form

$$\xi_t(i,j) = \frac{\alpha_t(i) P_{ij} \beta_{t+1}(j) b_j(y_{t+1})}{\sum_{i=1}^{K}\sum_{j=1}^{K} \alpha_t(i) P_{ij} \beta_{t+1}(j) b_j(y_{t+1})} \tag{3.14}$$

where $\alpha_t(i) = P[\mathbf{Y}^{(t)} = y^{(t)}, S_t = i; \lambda]$ $\mathbf{Y}^{(t)} = \{Y_1, ..., Y_t\}$,

$\beta_t(i) = P[\mathbf{Y}^{*(t)} = \mathbf{y}^{*(t)} \mid S_t = i; \lambda]$ $\mathbf{Y}^{*(t)} = \{Y_{t+1}, ..., Y_T\}$.

The second variable is defined as

$$\gamma_t(i) = P[S_t = i \mid \mathbf{Y} = \mathbf{y}, \lambda]$$

$$= \frac{P[S_t = i, \mathbf{Y} = \mathbf{y}; \lambda]}{P[\mathbf{Y} = \mathbf{y}; \lambda]}$$

$$= \frac{P[S_t = i, \mathbf{Y}^{(t)} = \mathbf{y}^{(t)}; \lambda] P[\mathbf{Y}^{*(t)} = \mathbf{y}^{*(t)} \mid S_t = i; \lambda]}{P[\mathbf{Y} = \mathbf{y}; \lambda]}, \tag{3.15}$$

which is the probability of being in state i at time t given the model and the observation sequence. This can be expressed in forward and backward variables by

$$\gamma_t(i) = \frac{\alpha_t(i)\beta_t(i)}{P[\mathbf{Y} = \mathbf{y}; \lambda]} = \frac{\alpha_t(i)\beta_t(i)}{\sum_{i=1}^{K} \alpha_t(i)\beta_t(i)} \tag{3.16}$$

and one can see that the relationship between $\gamma_t(i)$ and $\xi_t(i,j)$ is given by

$$\gamma_t(i) = \sum_{j=1}^{K} \xi_t(i,j), \qquad 1 \le i \le K, \qquad 1 \le t \le T. \tag{3.17}$$

If we sum $\gamma_t(i)$ over the time index t, we get a quantity, which can be interpreted as the expected (over time) number of times that state S_i is visited, or equivalently, the expected number of transitions made from state S_i (if we exclude the time slot $t = T$ from the summation). Similarly, summation of $\xi_t(i,j)$ over t (from $t = 1$ to $t = T - 1$)

can be interpreted as the expected number of transitions from state S_i to state S_j. That is

$$\sum_{t=1}^{T-1} \gamma_t(i) = \textit{Expected number of transition from } S_i \text{ and}$$

$$\sum_{t=1}^{T-1} \xi_t(i,j) = \textit{Expected number of transitions from } S_i \textit{ to } S_j .$$

Now assuming a starting model $\lambda = (A, B, \pi)$, we use the model to calculate the α's, β's using equations (3.3) to (3.8) then we use the α's and β's to calculate the ξ's and γ's using equations (3.14) to (3.17).

The next step is to define re-estimated model as $\hat{\lambda} = (\hat{A}, \hat{B}, \hat{\pi})$. The re-estimation formulas for $\hat{A}, \hat{B}, \hat{\pi}$ are

$$\hat{\pi}_i = \textit{Expected frequency in state } S_i \textit{ at time } t = 1$$

$$= \gamma_1(i), 1 \le i \le K . \tag{3.18}$$

$$\hat{P}_{ij} = \frac{\textit{Expected Number of transitions from state } S_i \textit{ to state } S_j}{\textit{Expected Number of transitions from state } S_i}$$

$$= \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)}, \qquad 1 \le i, j \le K . \tag{3.19}$$

$$\hat{b}_j(n) = \frac{\textit{Expected Number of times in state j and observing symbol } v_n}{\textit{Expected Number of times in state j}}$$

$$= \frac{\sum_{\substack{t=1 \\ s.t. O_t = v_n}}^{T} \gamma_t(j)}{\sum_{t=1}^{T} \gamma_t(j)}. \tag{3.20}$$

Equation (3.20) is in effect when the observations $\{Y_1, Y_2, ..., Y_T\}$ are discrete. The details about the continuous case are not mentioned here. We consider the discrete case throughout the thesis.

Suppose we have an initial guess of the parameters of the HMM $\lambda_0 = (A_0, B_0, \pi_0)$ and several sequences of observations. We can use these formulas to obtain a new model $\hat{\lambda}$ (i.e. the re-estimation model $\hat{\lambda} = (\hat{A}, \hat{B}, \hat{\pi})$), and it proves to be either of the following:

1. that the initial model λ defines a critical point of the likelihood function, in which case $\hat{\lambda} = \lambda$.
2. or, if $P[\mathbf{Y} = \mathbf{y}; \hat{\lambda}] > P[\mathbf{Y} = \mathbf{y}; \lambda]$, then it is the new model which best describes the observation sequence.

If we repeat these processes by using $\hat{\lambda}$ in place of λ, we can improve the probability of the observation sequence that is being produced by the model until the limiting point

is reached. The result of this procedure gives us the maximum likelihood estimator of the hidden Markov model (Baum et al., 1970).

It should be noted that the Baum-Welch method leads to a local maximum of λ only. In practice, to get a good solution, the initial guess λ_0 is very important. Usually several sets of starting guesses of λ_0 are used and one with the greatest likelihood value is chosen. Laird (1978) suggested a grid search method, which divides the searching domain into equally spaced small grids and starts from each of the intersections. Leroux and Puterman (1992) argue that the grid method would generate too many initial points when high dimensional spaces are involved and as such, they suggested a clustering algorithm.

CHAPTER 4

HIDDEN MARKOV MODEL AND THEIR APPLICATIONS TO WEED COUNTS

4.1 Introduction

Weed management makes a significant contribution to the harvesting of crops. Controlling weeds can improve the crop yield. It is also interesting to determine the relationship among more common weeds in the field. Another main factor of weed management is to find out whether there are different patterns or distributions within the field due to physical factors such as soil types, soil moisture and other reasons. The importance of these findings leads to better weed control practices.

In an agricultural survey conducted by Agriculture Canada, there were several fields considered without any treatments in Prairie Provinces. There were different kinds of weeds present in these fields, and some of the most common weeds found were Stinkweed, Wild Oats, Canada Thistle, Wild Buckwheat, Perennial Sow Thistle, Wild Mustard, Green Foxtail and Dandelion. In this thesis, one field has been selected (namely field #1; note that exact site location is not available), and the two most common weed types and one less frequent weed type are selected for analysis. The most

common species are Wild Oats (named as species 1509) and Wild Buckwheat (named as species 1097). Dandelion (named as species 1213) is less frequent. In the subsections of 4.2 give a description of each weed in the study. Figure 4.1, Figure 4.2, and Figure 4.3 and facts given in these sections are obtained from the Weed Identification Library (University of Manitoba, Department of Plant Science, 2005).

4.2 Weed species composition

There are differences in the species composition (richness and abundance) of the weed community depending on the type of fallow that preceded the cultivation phase. Note that there are some dominant species as well as abundant weed species present in these fields (Ngobo et al., 2004; Akobundu, 1999; and de Rouw, 1995). This dominant and abundant behaviour can be due to several reasons, such as species composition in fields and soil parameters, as well as some other environmental factors. The variation in weed species composition and abundance during the cropping phase is related to the physical and chemical characteristics of the soil. The physical and chemical soil properties are significantly correlated to the weed species composition (Ngobo et al., 2004). Unfortunately for this research, the background information of the fields is not available.

4.2.1 Wild Oats

Figure 4.1: Wild Oats

Wild Oats is an annual weed. Seeds can reproduce new plants. The seedling has a counter-clockwise leaf that forms into a spiral shape. The weed has hairs on the leaf margins. In the mature plant, the stems are smooth and vertical in position. This plant grows up to 150 cm tall. The leaves are not differing from Tame Oats (Figure 4.1). The head is an open panicle and the spikelets usually contain 2-3 florets (up to 7). The panicle may contain up to 200-250 seeds, ranging from black, brown, gray, yellow, to white.

4.2.1.1 Effects on crop quality

Wild Oats in grain is a reason for crop yield losses. Wild Oats compete for moisture, light, and nutrients. Barley and Canola are strong competitors, Wheat is an intermediate, and Oats and Flax are weak competitors.

Yield loss will depend on the number of Wild Oats per square metre and the stage of the Wild Oats and the crop. Left unchecked, 10 Wild Oat plants per square metre can reduce Wheat, Barley and Canola yields by 10% and Flax yields by 20%.

4.2.2 Wild Buckwheat

Figure 4.2: Wild Buckwheat

Wild Buckwheat is an annual weed that reproduces by seed. The stems are slightly angular, 30-90 cm long and freely branching at the base. The leaves are heart-shaped, pointed, 13-75 mm long, alternate, and smooth (Figure 4.2). The flowers are greenish-white and small. There are no petals but there are 5 sepals. Wild Buckwheat produces about 1,200 seeds per plant. It germinates under most soil conditions in cultivated fields and undeveloped areas.

4.2.2.1 Effects on crop quality

Wild Buckwheat makes swathing and combining difficult. With 5 plants per square metre, yield losses of 12% in Wheat can occur. With 30 plants per square metre, yield losses can jump to 22%. Yield losses of up to 10-20% have been reported in Flax with weed densities of 5-15 plants per square metre. The yield losses caused by this weed can be highly variable, depending on whether the weed emerges before, with, or after the crop.

4.2.3 Dandelion

Dandelions are present at all seasons of the year (perennials), and can be reproduced by seed. They are almost stemless and have deep thickset taproots. The leaves are in a rosette, 7.5-25 cm long and variable in shape (Figure 4.3). The flowers are bright yellow and are produced on a hollow, upright stem that is 30-45 cm in height. The seeds are 3 mm long and attached to a hairy parachute.

4.2.3.1 Effects on crop quality

Dandelion usually has little effect on forage crop quality. Established Dandelion causes yield losses in annual and winter annual crops and perennial forage seed crops. Dandelions can shorten the productive life of perennial forage seed crops. Dandelions also reduce productivity in pastures and hay crops.

Figure 4.3: Dandelion

4.3 Problem of interest and proposed solution

In an agricultural survey conducted by Agriculture Canada, there were several fields considered without any treatments in Prairie Provinces in Canada. There were different kinds of weeds present in these fields. The dataset used in this thesis was provided by Agriculture Canada to Prof. William H. Laverty, Mathematics and Statistics, University of Saskatchewan. As mentioned before (section 4.1), Wild Oats, Wild Buckwheat and Dandelion were selected for analysis. We assumed that these counts are multivariate Poisson variables which can be generated from multivariate Poisson distributions. The species counts are recorded from different fields and the fields are divided into an **a** x **b** grid. Weed species within $0.25m^2$ quadrates were identified and counted by species. For example, field #1 is divided into 10 x 15 grids (Figure 4.4). Four quadrats were assessed

at each of the 150 grid locations (Figure 4.5), with one quadrat sampled at a distance 1 meter north, south, east and west of the grid locations. Then take the total of the four quadrats as the weed count by species at that grid location. It is obvious that the observations recorded are spatially dependent. Also, one main goal of this study is to find out how many different clusters (states or distributions) are present in this field (Figure 4.6). The different clusters are formed due to factors such as the soil type, location and soil moisture or any other factor. Since only counts are recorded, the number of different clusters is unknown (i.e. hidden). Here, it can be assumed that this data structure follows a hidden Markov random field (HMRF).

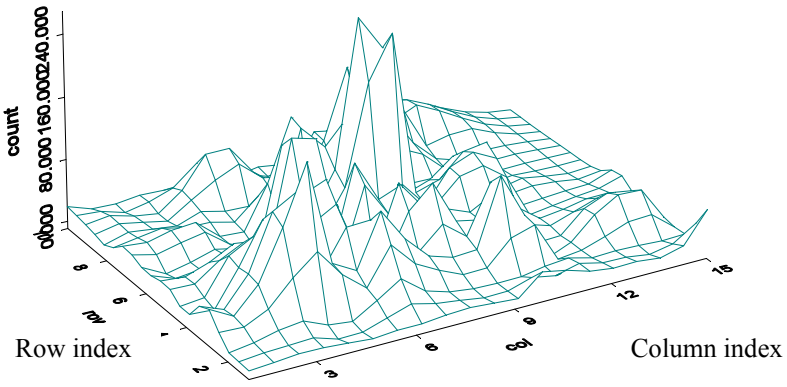


Figure 4.4: Distribution of weed counts in field #1

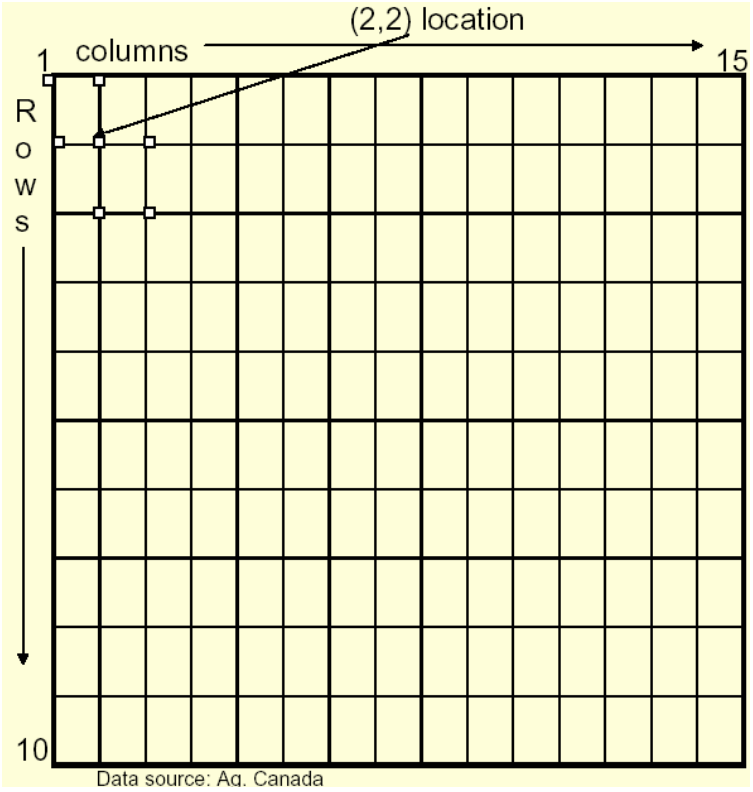


Figure 4.5: Data collection locations from field #1

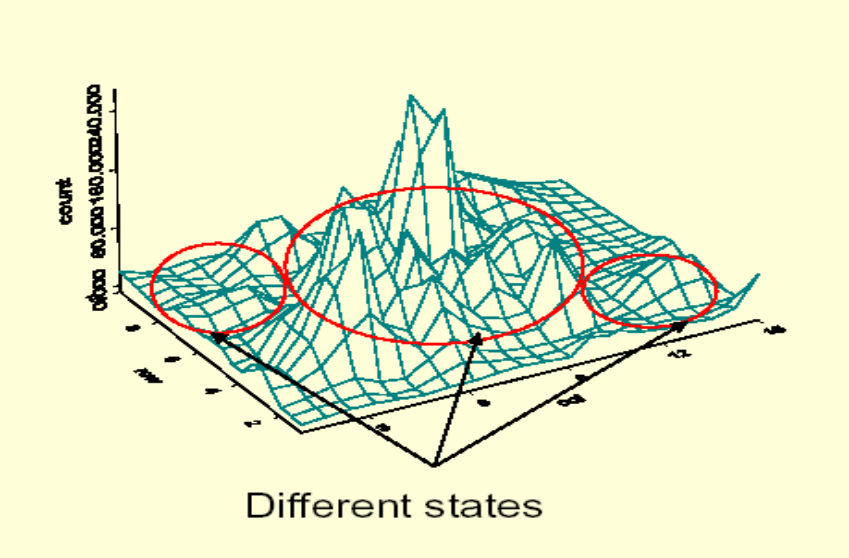


Figure 4.6: Distribution of weed counts and different states (clusters) in field #1

In the literature review (section 1.2.3), it is given that there is a very close relation with Markov random fields and Markov chains, and we explained that a 2-D (two-dimensional) Markov random field can be easily transformed into a 1-D (one-dimensional) Markov chain (Fjϕrtoft et al., 2003 and Aas et al., 1999). There are different types of scanning methods (Memon et al., 2000) available to transfer 2-D data into 1-D data (e.g. Line Scan, Hilbert-Peano scan etc.). The most common way is a scan line order, where the grid locations are traversed horizontally line by line (Figure 4.7). This method is called the line scan method (Memon et al., 2000; Bar-Joseph and Cohen-or, 2003).

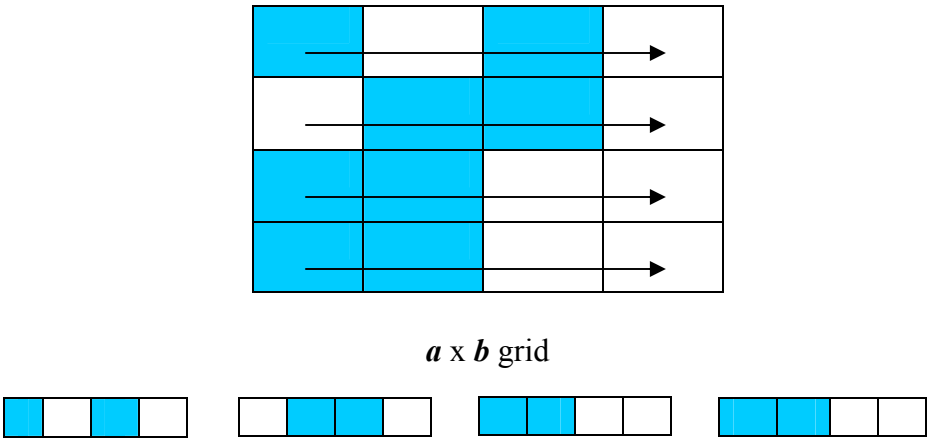


Figure 4.7: Scanning method: Line Scan

To illustrate how the line scan method can approximate spatial data, consider monthly (January-December) data across several years. This data can be arranged in a rectangular array (2-D) with rows representing years and columns representing months. Time series (1-D) periodic models exist where data from both neighbouring months in the same year and successive years in the same months are highly correlated. These 1-D

models would impose a spatial autocorrelation structure where both neighbouring E-W (East-West) and N-S (North-South) points would exhibit high correlation. A flaw in this approach is that the last point (December) in each row (year) is assumed to be highly correlated with the first point (January) in the consecutive row. This is not necessarily the case, however the effect of this assumption would be minimal. Thus the line scan method can transform the data into a 1-D sequence that is capable of approximating the spatial autocorrelation structure. If a periodic 1-D model is not used this would correspond to a spatial model where the neighbourhood system would only include E-W neighbours.

The weed species counts in the agriculture field also spatially correlated. The two-dimensional grid data can be transformed horizontally to a one-dimensional chain, by sweeping the $\boldsymbol{a}$ x $\boldsymbol{b}$ grid line by line. There will be slight irregularities in region borders with this approach rather than with the corresponding scheme based on Markov random field. However, the line scan transformation has less effect on irregularities since the agricultural field has a large neighbourhood system. That is, the distance between the neighbourhood points or coordinates in the agricultural field is large. In the literature review (section 1.2.3), it is given that the classification accuracy of the hidden Markov random field and the hidden Markov model will provide similar results and the hidden Markov model is much faster and simpler than the one based on Markov random fields (Fjϕrtoft et al., 2003 and Aas et al., 1999). Therefore, in this thesis, a novel multivariate Poisson hidden Markov model, which is a stochastic process, generated by a Markov chain whose state sequence cannot be observed directly but which can be indirectly

estimated through the observations is considered. As a comparison, finite mixture models are also discussed. The advantage of the multivariate Poisson hidden Markov model is that it takes serial correlation into account. Spatial information can be discovered by introducing a suitable covariance structure. To fit the multivariate Poisson hidden Markov model and the multivariate Poisson finite mixture model, the EM (expectation-maximization) algorithm is used.

4.4 Goals of the thesis:

1. Strategies for computation of multivariate Poisson probabilities:
    (i) Develop suitable recurrence relationships.
    (ii) Implement the recurrence relations using Splus/R software.
2. Univariate analysis for each species to find out how many clusters (or states) using finite mixture models and hidden Markov models.
3. Construct multivariate Poisson models with independent, common and restricted covariance structures and implement that in Splus/R software.
4. Fit a set of loglinear models for multivariate counts to decide the covariance structure.
5. Fit multivariate Poisson finite mixture models and multivariate Poisson hidden Markov models with independent, common and restricted covariance structures to determine the number of clusters.
6. Estimate the parameters of the distributions of clusters and calculate the standard errors using the bootstrap method.

7. Find out which observations are representative for each cluster.
8. Access the goodness of fit of the model using the entropy criterion, the unconditional covariance matrix and the information criterions.
9. Compare the two different methods in terms of the computational efficiency and the goodness of fit.
10. Usage of the model in the field of Agriculture.

In Chapter 2 and 3 details about Hidden Markov models are provided, with examples for the univariate case. The multivariate Poisson distribution, the calculation of probabilities and the multivariate Poisson finite mixture and the multivariate Poisson hidden Markov model estimation are discussed in Chapter 5.

CHAPTER 5

MULTIVARIATE POISSON DISTRIBUTION, MULTIVARIATE POISSON FINITE MIXTURE MODEL AND MULTIVARIATE POISSON HIDDEN MARKOV MODEL

5.1 The multivariate Poisson distribution: general description

Without loss of generality, the explanation in this thesis is restricted to three variables. Following the notation of Marshall and Olkin (1985), Johnson et al. (1997), Brijs (2002) and Brijs et al. (2004), the sets $R_1 = \{1,2,3\}$, $R_2 = \{12,13,23\}$, $R_3 = \{123\}$ are defined. Let $R = \bigcup_{i=1}^{3} R_i$. Now consider the independent variables X_j, which follow Poisson distributions with parameters θ_j with $j \in R$ respectively. Furthermore, the observed variables of interest Y_i, with $i = 1,2,3$ are defined as $Y_i = \sum_j X_j$ where $j \in R$ and j contains the subscript i. For example, the general, fully saturated covariance model for the case with three observed variables, where $R = \bigcup_{i=1}^{3} R_i$, is written as:

$$
\begin{aligned}
Y_1 &= X_1 + X_{12} + X_{13} + X_{123} \\
Y_2 &= X_2 + X_{12} + X_{23} + X_{123} \\
Y_3 &= X_3 + X_{13} + X_{23} + X_{123}.
\end{aligned}
$$

The parameters θ_j ($j \in R_m, m = 2,3$) correspond to a *m*-way covariance in a similar way to the *m*-way interactive terms, and thus, they impose structure on the data.

Mardia (1970) introduced the multivariate reduction technique to create the multivariate Poisson distribution. This reduction technique has been used extensively for the construction of multivariate models. The idea of the method is to start with some independent random variables and to create new variables by considering some functions of the original variables. Then, since the new variables contain jointly some of the original ones, a kind of structure is imposed creating multivariate models (Tsiamyrtzis et al., 2004).

We can represent this model using following matrix notations. Assume that X_i, $i = 1,...,k$ are independent Poisson random variables and **A** is a $n \times k$ matrix with zeros and ones. Then the vector $\mathbf{Y} = (Y_1, Y_2, ..., Y_n)$ defined as $\mathbf{Y} = \mathbf{AX}$ follows a n-variate Poisson distribution. The most general form assumes that **A** is a matrix of size $n \times (2^n - 1)$ of the form

$$\mathbf{A}=[\mathrm{A}_1, \mathrm{A}_2, \mathrm{A}_3,\ldots,\mathrm{A}_n]$$

where $\mathbf{A}_i$, $i = 1,...,n$ are matrices with n rows and $\binom{n}{i}$ columns. The matrix $\mathbf{A}_i$ contains columns with exactly i ones and $n - i$ zeros, with no duplicate columns, for $i = 1,...,n$. Thus $\mathbf{A}_n$ is the column vector of 1's while $\mathbf{A}_1$ becomes the identity matrix of size $n \times n$. For example, the fully structured multivariate Poisson model for three variables can be represented as follows:

$$Y_1 = X_1 + X_{12} + X_{13} + X_{123}$$
$$Y_2 = X_2 + X_{12} + X_{23} + X_{123}$$
$$Y_3 = X_3 + X_{13} + X_{23} + X_{123}$$

$$\mathbf{Y} = \mathbf{AX}$$

$$\mathbf{A}=[\mathrm{A}_1, \mathrm{A}_2, \mathrm{A}_3]$$

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & 1 \end{bmatrix}_{n \times k}$$

$$\mathbf{A}_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

where $\mathbf{A}_2 = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix}$

$$\mathbf{A}_3 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}.$$

Also we can write $\mathbf{Y} = \mathbf{AX}$ in more detailed as below.

$\mathbf{Y} = \mathbf{A}_1\mathbf{X}^{(1)} + \mathbf{A}_2\mathbf{X}^{(2)} + \mathbf{A}_3\mathbf{X}^{(3)}$ where

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \end{bmatrix}, \quad \mathbf{X}^{(1)} = \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix}, \quad \mathbf{X}^{(2)} = \begin{bmatrix} X_{12} \\ X_{13} \\ X_{23} \end{bmatrix}, \text{ and } \mathbf{X}^{(3)} = \begin{bmatrix} X_{123} \end{bmatrix}.$$

Moreover, this model enables us to construct some interesting submodels by appropriately defining the set R.

For example:

- If $R^* = R_1$ then the model reduces to an independence model (referred to as the local independence model).
- If $R^* = R_1 \cup R_3$ then the model reduces to a model with one common covariance term (referred to as the common covariance model).
- If the model assumes that $R^* = R_1 \cup R_2$ then it allows only two-way covariances (referred to as the restricted covariance model).

Note that omitting the set of parameters θ_j ($j \in R_m$), is equivalent to setting $\theta_j = 0$. The submodels can be formed by assuming that the corresponding θ's equal zero. Now, denote the cardinality of *R* as *J* which, for a trivariate model, equals *J* =7. Then, using the above notation, and considering the most general model with all the covariance terms (though it imposes unnecessarily large structure), the joint probability density of the corresponding multivariate Poisson distribution is given as

$$p(\mathbf{y} \mid \boldsymbol{\theta}) = P[Y_1 = y_1, Y_2 = y_2, Y_3 = y_3 \mid \theta_j, j \in R]$$
$$= \sum \cdots \sum \prod_{j \in R}^{J} Po(x_j \mid \theta_j),$$

where the summation is extended over all the combinations of x_j such that $y_i \geq \sum_k x_k$, $k \in R$ and k contains the subscript i. The fully-structured covariance model which is illustrated in section 5.1.1 needs four summations for the trivariate case, which obviously implies a large computational burden. The major problem of the use of the probability distribution in its general form is the calculation difficulty of the probability mass function. Kano and Kawamura (1991) described recursive schemes (section 5.2) to

reduce the computational burden; however the calculation remains computationally time-consuming for large dimensions.

This computational burden problem brings out the idea to create multivariate distributions with selected covariances, that is, not to include all the possible covariance terms, but only to select the covariance terms that are useful. In reality, using all the *m*-fold covariance terms imposes too much structure, while complicating the whole procedure without adding any further insight into the data. For this reason, after a preliminary assessment, one may identify interesting covariance terms that may be included into the model. This selection corresponds to fixing the value of the Poisson parameters, that is, the corresponding θ's.

Based on this general description of the multivariate Poisson distribution and the relationship with more suitable submodels, a detailed description of each model is provided in the next few sections.

5.1.1 The fully- structured multivariate Poisson model

The theoretical development of the fully structured multivariate Poisson model will be illustrated by weed species: Wild Buckwheat, Dandelion and Wild Oats. Suppose the objective is to cluster weed count data based on the mean counts in a set of three weed species, that is, Wild Buckwheat (Y_1), Dandelion (Y_2), Wild Oats (Y_3). Following the

notation of Marshall and Olkin (1985), and Johnson et al. (1997), Brijs (2002) and Brijs et al. (2004), and based on the discussion in section 5.1, a trivariate Poisson random variable (Y_1, Y_2, Y_3) with parameters $(\theta_1, \theta_2, \theta_3, \theta_{12}, \theta_{13}, \theta_{23}, \theta_{123})$ can then be constructed from a number of independent univariate Poisson distributions as follows:

$$Y_1 = X_1 + X_{12} + X_{13} + X_{123}$$
$$Y_2 = X_2 + X_{12} + X_{23} + X_{123}$$
$$Y_3 = X_3 + X_{13} + X_{23} + X_{123}$$

with all X's are independent univariate Poisson distributions with their respective means $\theta_1,\ \theta_2,\ \theta_3,\ \theta_{12},\ \theta_{13},\ \theta_{23},\ \theta_{123}$. The calculation of the probability distribution of $P[Y_1 = y_1, Y_2 = y_2, Y_3 = y_3]$ is not easy. The solution is based on the observation that $P[Y_1 = y_1, Y_2 = y_2, Y_3 = y_3]$ is the marginal distribution from $P[Y_1 = y_1, Y_2 = y_2, Y_3 = y_3, X_{12} = x_{12}, X_{13} = x_{13}, X_{23} = x_{23}, X_{123} = x_{123}]$ and can be obtained by summing out over all X's, i.e.,

$$P[Y_1 = y_1, Y_2 = y_2, Y_3 = y_3] = \sum_{x_{12}=0}^{L_1} \sum_{x_{13}=0}^{L_2} \sum_{x_{23}=0}^{L_3} \sum_{x_{123}=0}^{L_4} P[Y_1 = y_1, Y_2 = y_2, Y_3 = y_3, X_{12} = x_{12}, X_{13} = x_{13}, X_{23} = x_{23}, X_{123} = x_{123}] \quad (5.1)$$

with $L_1 = \min(y_1, y_2)$,

$L_2 = \min(y_1 - x_{12}, y_3)$,

$L_3 = \min(y_2 - x_{12}, y_3 - x_{13})$,

$L_4 = \min(y_1 - x_{12} - x_{13}, y_2 - x_{12} - x_{23}, y_3 - x_{13} - x_{23})$.

The above expression (5.1) demonstrates that the *x*'s are summed out over all possible values of the respective *X*'s. It is known that the *X*'s should take only positive integer values or zero, since the *X*'s are Poisson distributed variables. However, the upper bounds (*L's*) of the different *X*'s are unknown and will depend on the values of $y_1,\ y_2$,

and y_3 as illustrated above. Again, to rewrite equation 5.1 in compact form, we can define the following notations:

$$\mathbf{L}^{(2)} = \begin{bmatrix} L_1 \\ L_2 \\ L_3 \end{bmatrix}, \quad \mathbf{L}^{(3)} = \begin{bmatrix} L_4 \end{bmatrix} \quad \text{and} \quad R_2 = (R_2^{(1)} \cup R_2^{(2)} \cup R_2^{(3)}) \quad \text{where} \quad R_2^{(1)} = \{12,13\},$$

$R_2^{(2)} = \{12,23\}$, $R_2^{(3)} = \{13,23\}$.

In fact, substituting the X's for the Y's in (5.1) result in:

$$P[Y_1 = y_1, Y_2 = y_2, Y_3 = y_3] = \sum_{\mathbf{x}^{(2)}=0}^{\mathbf{L}^{(2)}} \sum_{\mathbf{x}^{(3)}=0}^{\mathbf{L}^{(3)}} P[X_1 = x_1, X_2 = x_2, X_3 = x_3, X_{12} = x_{12}, X_{13} = x_{13}, X_{23} = x_{23}, X_{123} = x_{123}]$$

and since the X's are independent univariate Poisson variables, the joint distribution reduces to the product of the univariate distributions:

$$P[Y_1 = y_1, Y_2 = y_2, Y_3 = y_3] = \sum_{\mathbf{x}^{(2)}=0}^{\mathbf{L}^{(2)}} \sum_{\mathbf{x}^{(3)}=0}^{\mathbf{L}^{(3)}} \prod_{j \in R_1} P[X_j = (y_j - \sum_{k \in R_2^{(j)}} x_k - \sum_{l \in R_3} x_l)] \prod_{i \in R_2 \cup R_3} P[X_i = x_i].$$

Now, the following three conditions on X_1, X_2, and X_3 must be satisfied, since the X's are univariate Poisson distributions, and since the Poisson distribution is only defined for positive integer values and zero:

$$\begin{aligned} y_1 - x_{12} - x_{13} - x_{123} &\geq 0 \\ y_2 - x_{12} - x_{23} - x_{123} &\geq 0 \\ y_3 - x_{13} - x_{23} - x_{123} &\geq 0. \end{aligned} \tag{5.2}$$

These conditions imply that all x's cannot just be any integer value, but depend on the values of y_1, y_2, and y_3. Accordingly, the distribution for $P[Y_1 = y_1, Y_2 = y_2, Y_3 = y_3]$ by summing up all the x's is:

$$P[Y_1 = y_1, Y_2 = y_2, Y_3 = y_3] = \sum_{\mathbf{x}^{(2)}=0}^{\mathbf{L}^{(2)}} \sum_{\mathbf{x}^{(3)}=0}^{\mathbf{L}^{(3)}} e^{-\theta} \frac{\prod_{j \in R_1} \theta_j^{(y_j - \sum_{k \in R_2^{(j)}} x_k - \sum_{l \in R_3} x_l)} \prod_{i \in R_2 \cup R_3} \theta_i^{x_i}}{\prod_{j \in R_1} (y_j - \sum_{k \in R_2^{(j)}} x_k - \sum_{l \in R_3} x_l)! \prod_{i \in R_2 \cup R_3} x_i!}$$

with $\theta = \theta_1 + \theta_2 + \theta_3 + \theta_{12} + \theta_{13} + \theta_{23} + \theta_{123}$.

To see why L_1 must be equal to $\min(y_1, y_2)$, one must look at the first two of the three conditions on X_1, X_2, and X_3 specified in (5.2). Indeed, it is known that all x's should be zero or a positive integer. For that reason, if all x's except for x_{12} would be zero, then the maximum acceptable value for the x_{12} can be $\min(y_1, y_2)$ to facilitate the first two conditions. Similarly, the values for the other x's can be computed, based on the preceding values (Mahamunulu, 1967), resulting in the admissible ranges for all *L*'s.

The above formulated trivariate Poisson model incorporates all possible interactions (that is, two-way and three-way) that can exist between the counts of the three weed species considered. In other words, this model can take into account for all possible covariances between the weed counts.

The mixture variant of the multivariate Poisson model (details are given in section 5.3) simply extends the multivariate Poisson model by assuming that *k* groups of species have different parameter values for the θ's. Clearly, the number of parameters to be optimized rapidly increase with the specification of different groups in the data.

In general, the number of parameters to be estimated is equal to $(k-1)+k\times(2^q-1)$, for q varaites k-component mixture model. The number of parameters to be estimated increases linearly in the number of components (k) and exponentially in the number of the variables (q) being considered.

5.1.2 The multivariate Poisson model with common covariance structure

The fully structured multivariate Poisson model (section 5.1.1) has a large number of parameters that need to be estimated. Therefore, an alternative approach has been proposed in the literature to make a simpler version of the model by representing variance/covariance by means of one common term (Johnson and Kotz, 1969, Li et al., 1999 and Karlis, 2003).

In this approach, following the explanation in section 5.1, the trivariate Poisson variable (Y_1, Y_2, Y_3) with one common covariance term, is defined as the following:

$$Y_1 = X_1 + X_{123}$$
$$Y_2 = X_2 + X_{123}$$
$$Y_3 = X_3 + X_{123}$$

with all X's independent univariate Poisson distribution with respective parameters θ_1, θ_2, θ_3, and θ_{123}.

In a matrix notation, this model can be presented as:

$$\mathbf{Y} = \mathbf{AX}$$

$$\mathbf{A}=[\mathrm{A}_1, \mathrm{A}_3]$$

$$\mathbf{A}=\begin{bmatrix}1 & 0 & 0 & 1\\ 0 & 1 & 0 & 1\\ 0 & 0 & 1 & 1\end{bmatrix}$$

where $\mathbf{A}_1=\begin{bmatrix}1 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & 1\end{bmatrix}$

$\mathbf{A}_3=\begin{bmatrix}1\\ 1\\ 1\end{bmatrix}$.

Although the covariance structure is limited compared to the general definition of the multivariate Poisson, there are a convenient number of parameters to deal with. Moreover, trivariate Poisson distribution $P[Y_1=y_1, Y_2=y_2, Y_3=y_3]$ can now be obtained as the marginal distribution from $P[Y_1=y_1, Y_2=y_2, Y_3=y_2, X_{123}=x_{123}]$ as follows:

$$P[Y_1=y_1, Y_2=y_2, Y_3=y_3]=\sum_{\mathbf{x}^{(3)}=0}^{\min(y_1,y_2,y_3)} P[Y_1=y_1, Y_2=y_2, Y_3=y_3, X_{123}=x_{123}]. \qquad (5.3)$$

Substituting the X's for the Y's in (5.3) results in:

$P[Y_1=y_1, Y_2=y_2, Y_3=y_3]=\sum_{\mathbf{x}^{(3)}=0}^{\mathbf{L}^{(4)}} P[X_1=x_1, X_2=x_2, X_3=x_3, X_{123}=x_{123}]$ with all X's independent univariate Poisson distributions and $\mathbf{L}^{(4)}=\left[L_5\right]$ where $L_5=\min(y_1, y_2, y_3)$, thus:

$$P[Y_1=y_1, Y_2=y_2, Y_3=y_3]=\sum_{\mathbf{x}^{(3)}=0}^{\mathbf{L}^{(4)}}\prod_{j\in R_1} P[X_j=(y_j-\sum_{l\in R_3} x_l)]\prod_{i\in R_3} P[X_i=x_i]$$

$$=\sum_{\mathbf{x}^{(3)}=0}^{\mathbf{L}^{(4)}} e^{-(\theta_1+\theta_2+\theta_3+\theta_{123})}\frac{\prod_{j\in R_1}\theta_j^{(y_j-\sum_{l\in R_3} x_l)}\prod_{i\in R_3}\theta_i^{x_i}}{\prod_{j\in R_1}(y_j-\sum_{l\in R_3} x_l)!\prod_{i\in R_3} x_i!}.$$

Similar to the fully structured model presented in the previous section, the general *k*-components *q*-variate Poisson mixture model requires the estimation of $(k-1)+k\times(q+1)$ parameters. The number of parameters increases linearly both with the number of the variables and components being considered.

5.1.3 The multivariate Poisson model with local independence

In the previous sections, it was revealed that in the case of three weed species, the joint probability of observing multiple outcomes $P[Y_{1i}=y_{1i},Y_{2i}=y_{2i},Y_{3i}=y_{3i}]$ for an 'i'$^{\text{th}}$ location is distributed according to the multivariate Poisson distribution. On the other hand, under the assumption of local independence of the weed count rates within each mixture component, this joint probability reduces to the product of the weed species-specific densities, that is,

$$P[Y_{1i}=y_{1i},Y_{2i}=y_{2i},Y_{3i}=y_{3i}]=P[Y_{1i}=y_{1i}]\times P[Y_{2i}=y_{2i}]\times P[Y_{3i}=y_{3i}].$$

This means that the following representation is obtained for the Y_i's:

$$Y_1=X_1$$
$$Y_2=X_2$$
$$Y_3=X_3.$$

In matrix notations, this model can be presented as:

$$\mathbf{Y}=\mathbf{AX}$$

$$\mathbf{A}=[A_1]$$

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

where $\mathbf{A}_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$.

In this situation, the likelihood function for the general k component mixture model for q weed species takes a very simple form:

$L(\Theta; y_{li}) = \prod_{i=1}^{n} f(y_{li} \mid \Theta) = \prod_{i=1}^{n} \sum_{j=1}^{k} p_j \prod_{l=1}^{q} \frac{(\boldsymbol{\theta}_{lj})^{y_{il}} \exp(-\boldsymbol{\theta}_{lj})}{y_{li}!}$, where p_j are mixing proportions.

Moreover, for the general k component mixture model for q weed species, we have $k-1$ different p's, and k different θ's per weed species. The number of parameters need to be estimated is $(k-1)+k\times q$. Details of the multivariate Poisson finite mixture model is given in section 5.3. The loglikelihood is then expressed as:

$$LL(\forall p, \boldsymbol{\theta} \mid data) = \sum_{i=1}^{n} \ln\left[\sum_{j=1}^{k} p_j \prod_{l=1}^{q} \frac{(\boldsymbol{\theta}_{lj})^{y_{li}} \exp(-\boldsymbol{\theta}_{lj})}{y_{li}!} \right].$$

5.1.4 The multivariate Poisson model with restricted covariance

The multivariate Poisson models presented in section 5.1.1 and 5.1.2 represent two extreme approaches to model the interdependent count rates. From a theoretical aspect, the fully structured model is preferable to the model with common covariance structure

because the former captures more of the existing variance in the data. However, from a practical aspect, the model with common covariance structure is preferable to the fully structured model because it requires fewer parameters to be estimated.

Therefore, the important question is whether a model somewhere in between the two presented extreme models can be found, that is, both a) theoretically good enough to describe most of the existing covariances, and b) practically suitable in terms of the number of parameters to be estimated.

The model introduced in this section was tried to address the above mentioned problem. The main idea is to simplify the variance/covariance structure as much as possible by including only statistically significant *m*-fold interactions. For this trivariate model, the statistical significance of the weed count interactions between Wild Buckwheat (Y_1), Dandelion (Y_2) and Wild Oats (Y_3) will study by using of loglinear analysis (see section 5.7 and section 6.3). The loglinear analysis is particularly appropriate for the development of a simpler multivariate Poisson mixture model for clustering since it facilitates to discover, which interaction terms in the variance/covariance matrix can be set equal to zero. We will show that (section 6.3) the p values of the goodness of fit statistics of the models with some of the two-fold interactions and without any interaction do not differ very much. Therefore, the two-fold interactions were kept in the model. The latent variables, $X = (X_1, X_2, X_3, X_{12}, X_{13}, X_{23})$ and the vector of parameters, $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3, \theta_{12}, \theta_{13}, \theta_{23})$, are used to present the model and thus the restricted covariance model can be defined as:

$$Y_1 = X_1 + X_{12} + X_{13}$$
$$Y_2 = X_2 + X_{12} + X_{23}$$
$$Y_3 = X_3 + X_{13} + X_{23}.$$

The joint probability function of an observation $\mathbf{Y} = (Y_1, Y_2, Y_3)$ is given (Karlis, 2004) as

$$p(\mathbf{y};\boldsymbol{\theta}) = P[Y_1 = y_1, Y_2 = y_2, Y_3 = y_3] = \sum_{\mathbf{x}^{(3)}}^{\mathbf{L}^{(2)}} \exp(-\sum_{m\in\mathbf{A}} \theta_m) \frac{\prod_{j\in R_1} \theta_j^{(y_j - \sum_{k\in R_2^{(j)}} x_k)} \prod_{i\in R_2} \theta_i^{x_i}}{\prod_{j\in R_1} (y_j - \sum_{k\in R_2^{(j)}} x_k)! \prod_{i\in R_2} x_i!} \ ,$$

where $\mathbf{A} = \{1,2,3,12,13,23\}$. The unconditional probability mass function is given under a mixture with k-components model by $\sum_{j=1}^{k} p_j p(\mathbf{y};\boldsymbol{\theta}_j) = \sum_{j=1}^{k} p_j p(y_1, y_2, y_3; \boldsymbol{\theta}_j)$.

As a result, the model assumes covariance between all the variables since it is imposed by the mixing distribution. For a model with k components the number of parameters equals to $7k$-1 (that is, 6 theta's per component plus the mixing proportions), which, compared to the fully saturated model that contains $(k-1)+k\times(2^q-1)$ parameters, increases linearly instead of exponentially with the number of components considered.

5.2 Computation of multivariate Poisson probabilities

The multivariate Poisson distribution is one of the well-known and important multivariate discrete distributions. Nevertheless this distribution has not found a lot of practical applications except the special case of the bivariate Poisson distribution

(Tsiamyrtzis and Karlis, 2004). The main reason for this is the unmanageable probability function, which causes inferential procedures to be somewhat problematic and difficult. For example, consider the estimation of maximum likelihood (ML) estimates. To estimate the likelihood, one has to calculate the probability function at all the observations. The probability function can be calculated via recurrence relationships, otherwise exhausting summations are needed (Tsiamyrtzis and Karlis, 2004). The efficient algorithm must be used to do the calculation of probabilities (especially for higher dimensions) to save time. Applying purely the recurrence relationships without trying to use them in an optimal way can be difficult and time-consuming (Tsiamyrtzis and Karlis, 2004). For further motivation, consider a problem with, say three-dimensional data. For example, the data may represent the three different weed species counts in a field. If the number of locations is not very large, this implies that it will result many cells with zero frequency, so the calculation of the entire three-dimensional space of all combinations for the number of count (plants) of the three weed species is a very bad approach. For instance, if the maximum number of counts for each species is denoted as a_i, then to create the entire probability table, one has to calculate $\prod_{i=1}^{3}(a_i+1)$ different probabilities using normal strategy. Clearly, the calculation of these probabilities is awkward and time-consuming. It is especially usual if one can only calculate the non-zero frequency cells, which contribute to the likelihood. Tsiamyrtzis and Karlis (2004) proposed efficient strategies for calculating the multivariate Poisson probabilities based on the existing recurrence relationships.

5.2.1 The multivariate Poisson distribution with common covariance

Suppose that X_i are independent Poisson random variables with mean θ_i, for $i = 0,1,...,n$ and let $Y_i = X_0 + X_i, i = 1,...,n$. Then the random vector $\mathbf{Y} = (Y_1, Y_2, ..., Y_n)$ follows a *n*-variate Poisson distribution, where *n* denotes the dimension of the distribution. The joint probability function (Karlis, 2003) is given by

$$p(\mathbf{y};\boldsymbol{\theta}) = P[Y_1 = y_1, Y_2 = y_2, ..., Y_n = y_n]$$

$$= \exp\left(-\sum_{i=1}^{n}\theta_i\right)\prod_{i=1}^{n}\frac{\theta_i^{y_i}}{y_i!}\sum_{i=0}^{s}\left[\prod_{j=1}^{n}\binom{y_j}{i} i!\left(\frac{\theta_0}{\prod_{k=1}^{n}\theta_k}\right)^i\right], \qquad (5.4)$$

where $s = \min\{y_1, y_2, ..., y_n\}$. Marginally each Y_i follows a Poisson distribution with parameter $\theta_0 + \theta_i$. The parameter θ_0 (common covariance) is the covariance between all the pairs of random variables (Y_i, Y_j) where $i, j \in \{1,...,n\}$ and $i \neq j$. If $\theta_0 = 0$ then the variables are independent and the multivariate Poisson distribution reduces to the product of independent Poisson distributions. The recurrence relations can be applied to compute the above probabilities. A general scheme for constructing recurrence relations for multivariate Poisson distributions was provided by Kano and Kawamura (1991). Details are given below.

Some notations are introduced first to make it easy to explain the distributions. Let **0** and **1** denote the vector with all elements equal to 0 and 1 respectively and $\mathbf{e_i}$ the unit vector with all elements 0 except from the i^{th} element which is equal to 1. Using this

notation, the following recursive scheme is proved for the multivariate Poisson distribution presented in (5.4):

$$y_i p(\mathbf{y}) = \theta_i p(\mathbf{y} - \mathbf{e_i}) + \theta_0 p(\mathbf{y} - \mathbf{1}), \;\; i = 1,...,n \tag{5.5}$$

$$P[Y_1 = y_1,...,Y_k = y_k, Y_{k+1} = 0,...,Y_n = 0] = p\left(\mathbf{y} - \sum_{i=1}^{k} \mathbf{e_i}\right) \prod_{i=1}^{k} \frac{\theta_i}{y_i}, \quad \text{for } k = 1,...,n-1, \tag{5.6}$$

where the order of Y_i's and 0's can be interchanged to cover all possible cases, while $p(\mathbf{0}) = \exp(-\sum_{i=0}^{n} \theta_i)$. The recurrence equation (5.6) holds for arbitary permutations of Y_i's.

It can be seen that since at every case at least one of the y_i's equals 0, i.e. $s = 0$, the sum appearing in the joint probability function has just one term and hence the joint probability function takes the useful form $P[\mathbf{Y} = \mathbf{y}] = \exp(-\theta_0) \prod_{i=1}^{n} Po(y_i; \theta_i)$, where $Po(y; \theta) = \exp(-\theta) \frac{\theta^y}{y!}$ denotes the probability function of the simple Poisson distribution with a parameter θ. Then equation (5.6) arises by using the recurrence relation for the univariate Poisson distribution (Tsiamyrtzis and Karlis, 2004).

Two examples of recurrence relations for the common covariance multivariate Poisson distribution are given below. The $\boldsymbol{\theta}$ in all recurrence relations is suppressed for simplicity of the notation.

The bivariate Poisson distribution has joint probability function given by:

$$p(y_1, y_2) = P[Y_1 = y_1, Y_2 = y_2] = e^{-(\theta_0 + \theta_1 + \theta_2)} \frac{\theta_1^{y_1}}{y_1!} \frac{\theta_2^{y_2}}{y_2!} \sum_{i=0}^{s} \binom{y_1}{i} \binom{y_2}{i} i! \left(\frac{\theta_0}{\theta_1 \theta_2}\right)^i,$$

where $s = \min\{y_1, y_2\}$. According to the general recurrence in (5.5) the following recurrence are found:

$$y_1 p(y_1, y_2) = \theta_1 p(y_1 - 1, y_2) + \theta_0 p(y_1 - 1, y_2 - 1)$$
$$y_2 p(y_1, y_2) = \theta_2 p(y_1, y_2 - 1) + \theta_0 p(y_1 - 1, y_2 - 1),$$

with the convention that $p(y_1, y_2) = 0,$ if $s < 0.$ Using these two recurrence relationships interchangeably, one can get the entire probability table with $\prod_{i=1}^{2}(y_i + 1)$ probabilities.

The trivariate Poisson distribution has joint probability function given by following:

$$p(y_1, y_2, y_3) = P[Y_1 = y_1, Y_2 = y_2, Y_3 = y_3] = e^{-(\theta_0+\theta_1+\theta_2+\theta_3)} \frac{\theta_1^{y_1}}{y_1!} \frac{\theta_2^{y_2}}{y_2!} \frac{\theta_3^{y_3}}{y_3!} \sum_{i=0}^{s} \binom{y_1}{i} \binom{y_2}{i} \binom{y_3}{i} i! \left(\frac{\theta_0}{\theta_1 \theta_2 \theta_3} \right)^i,$$

where $s = \min\{y_1, y_2, y_3\}$. Using the general recurrence in (5.5) the following recurrence are found:

$$y_1 p(y_1, y_2, y_3) = \theta_1 p(y_1 - 1, y_2, y_3) + \theta_0 p(y_1 - 1, y_2 - 1, y_3 - 1)$$
$$y_2 p(y_1, y_2, y_3) = \theta_2 p(y_1, y_2 - 1, y_3) + \theta_0 p(y_1 - 1, y_2 - 1, y_3 - 1)$$
$$y_3 p(y_1, y_2, y_3) = \theta_3 p(y_1, y_2, y_3 - 1) + \theta_0 p(y_1 - 1, y_2 - 1, y_3 - 1)$$

with the convention that $p(y_1, y_2, y_3) = 0,$ if $s < 0.$

Tsiamyrtzis and Karlis (2004) demonstrated that how to use these existing recurrence relationships efficiently to calculate probabilities using two algorithms called the Flat and Full algorithms. This proposed algorithm can be extended to a more general multivariate Poisson distribution that allows full structure with terms for all the pairwise covariances, covariance among three variables and so on (Mahamunulu, 1967).

5.2.2 The multivariate Poisson distribution with restricted covariance

The fully structured multivariate Poisson model has not found any real data applications because of the complicated form of the probability function and the excessive structure of the model. In this thesis, the restricted covariance structure, which explains only pairwise covariances, is considered. The restricted covariance multivariate Poisson model can be presented as follows:

$$Y_1 = X_1 + X_{12} + X_{13}$$
$$Y_2 = X_2 + X_{12} + X_{23}$$
$$Y_3 = X_3 + X_{13} + X_{23}$$

$$\mathbf{A}=[\mathrm{A}_1, \mathrm{A}_2]$$

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 \end{bmatrix}$$

where

$$\mathbf{A}_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$
$$\mathbf{A}_2 = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix}.$$

Details about the multivariate Poisson model with two-way covariance structure can also be found in Karlis and Meligkotsidou (2005).

Marginal distributions are Poisson:

$$Y_i \sim Poisson(\theta_i + \theta_{ij} + \theta_{ik}),\ \begin{cases} i = 1,...,n \\ j,k = 2,...,n. \\ j,k \neq i \end{cases}$$

The joint probability function is given by

$$p(y_1,y_2,y_3;\boldsymbol{\theta}) = P[Y_1 = y_1, Y_2 = y_2, Y_3 = y_3] = \sum_{\mathbf{x}^{(3)}}^{\mathbf{L}^{(2)}} \exp(-\sum_{m\in\mathbf{A}} \theta_m) \frac{\prod_{j\in R_1} \theta_j^{(y_j - \sum_{k\in R_2^{(j)}} x_k)} \prod_{i\in R_2} \theta_i^{x_i}}{\prod_{j\in R_1}(y_j - \sum_{k\in R_2^{(j)}} x_k)! \prod_{i\in R_2} x_i!} \quad .$$

The calculation of above joint probability function is not easy. Here, we used recurrence relations involving densities to compute the probability function. In 1967, Mahamunulu presented some important notes regarding p variate Poisson distributions. According to him the following recurrence relations are obtained for the trivariate two-way covariance model:

$$y_1 p(y_1,y_2,y_3) = \theta_1 p(y_1-1,y_2,y_3) + \theta_{12} p(y_1-1,y_2-1,y_3) + \theta_{13} p(y_1-1,y_2,y_3-1) \tag{5.7}$$

$$y_2 p(y_1,y_2,y_3) = \theta_2 p(y_1,y_2-1,y_3) + \theta_{12} p(y_1-1,y_2-1,y_3) + \theta_{23} p(y_1,y_2-1,y_3-1) \tag{5.8}$$

$$y_3 p(y_1,y_2,y_3) = \theta_3 p(y_1,y_2,y_3-1) + \theta_{13} p(y_1-1,y_2,y_3-1) + \theta_{23} p(y_1,y_2-1,y_3-1), \tag{5.9}$$

with $p(y_1,y_2,y_3) = 0$ if $\min\{y_1,y_2,y_3\} < 0$.

It also gives the following relations:

$$\left.\begin{aligned} y_1 p(y_1,y_2,0) &= \theta_1 p(y_1-1,y_2,0) + \theta_{12} p(y_1-1,y_2-1,0) \quad y_1, y_2 \geq 1 \\ y_2 p(0,y_2,y_3) &= \theta_2 p(0,y_2-1,y_3) + \theta_{23} p(0,y_2-1,y_3-1) \quad y_2, y_3 \geq 1 \\ y_3 p(y_1,0,y_3) &= \theta_3 p(y_1,0,y_3-1) + \theta_{13} p(y_1-1,0,y_3-1) \quad y_1, y_3 \geq 1 \end{aligned}\right\} \tag{5.10}$$

$$\left.\begin{aligned} y_1 p(y_1,0,0) &= \theta_1 p(y_1-1,0,0) \qquad y_1 \geq 1 \\ y_2 p(0,y_2,0) &= \theta_2 p(0,y_2-1,0) \qquad y_2 \geq 1 \\ y_3 p(0,0,y_3) &= \theta_3 p(0,0,y_3-1) \qquad y_3 \geq 1 \end{aligned}\right\} \tag{5.11}$$

$$p(0,0,0) = \exp(-(\theta_1+\theta_2+\theta_3+\theta_{12}+\theta_{13}+\theta_{23})). \tag{5.12}$$

The above mentioned recurrence relations and the Flat algorithm (Tsiamyrtzis and Karlis, 2004) are used to calculate the probabilities of the restricted covariance trivariate Poisson model.

5.2.3 The Flat algorithm

Using the Flat Algorithm the calculation of $p(y_1,y_2,y_3)$ can be done in two stages. In the first stage, one can move from (y_1,y_2,y_3) to the closest hyperplane using only one of the recurrence relationships (5.7), and in the second stage, he can move down to the **0** point by the simplified recurrence relationships (5.10) and (5.11). Thus, starting from (y_1,y_2,y_3) and applying the recurrence relationship, we get three new points (y_1-1,y_2,y_3), (y_1-1,y_2-1,y_3) and (y_1-1,y_2-1,y_3-1). Applying the same recurrence relationship to these three points we get another six new points: (y_1-2,y_2,y_3), (y_1-2,y_2-1,y_3), (y_1-2,y_2,y_3-1), (y_1-2,y_2-2,y_3), (y_1-2,y_2-1,y_3-1) and (y_1-2,y_2,y_3-2). Figure 5.1 illustrates how coordinates can move to the closer plane using the recurrence relationship (5.7) for the case $y_1 \leq y_2 \leq y_3$.

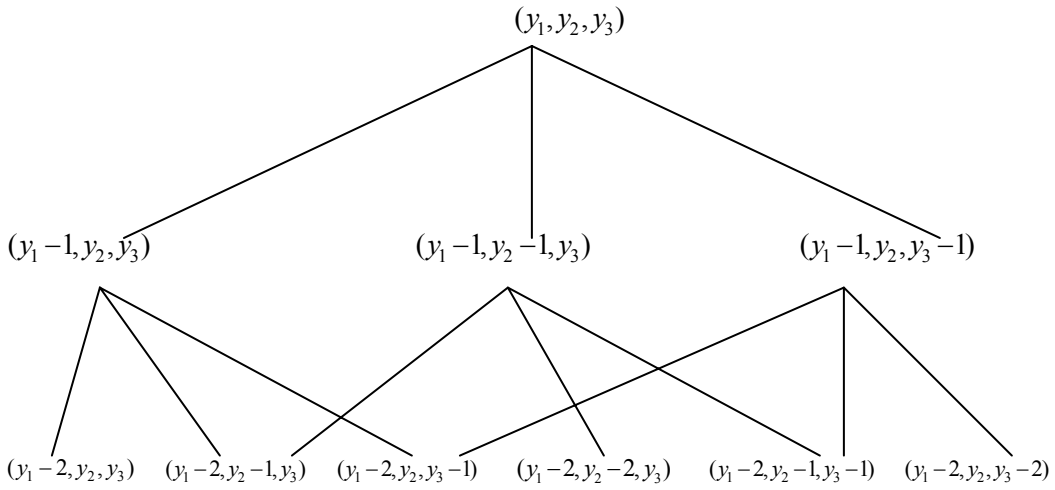


Figure 5.1: Flat algorithm (stage 1)

Using the Flat algorithm, one can move along a plane until the minimum coordinate equal to zero (stage 2). For example, consider the calculation of probability $p(2,2,2)$. Figure 5.2 and Figure 5.4 illustrate how the Flat algorithm works.

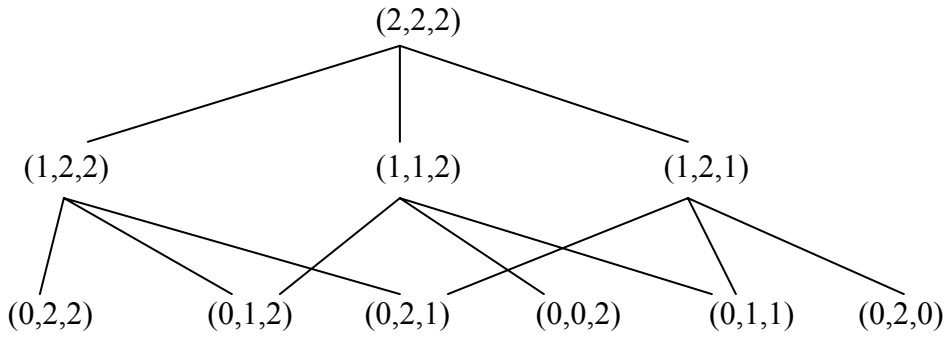


Figure 5.2: Calculating $p(2,2,2)$ using the Flat algorithm

When you come to this stage, you can use the Flat algorithm for $p(y_1, y_2)$ or $p(y_1, y_3)$ or $p(y_2, y_3)$ according to Figure 5.3. Thus, starting from (y_1, y_2, y_3) and applying the

recurrence relationship $y_1 p(y_1, y_2, 0) = \theta_1 p(y_1 - 1, y_2, 0) + \theta_{12} p(y_1 - 1, y_2 - 1, 0)$ we get two new points $(y_1 - 1, y_2)$ and $(y_1 - 1, y_2 - 1)$. Applying the same recurrence relationship to these two points, we get another three new points: $(y_1 - 2, y_2)$, $(y_1 - 2, y_2 - 1)$ and $(y_1 - 2, y_2 - 2)$.

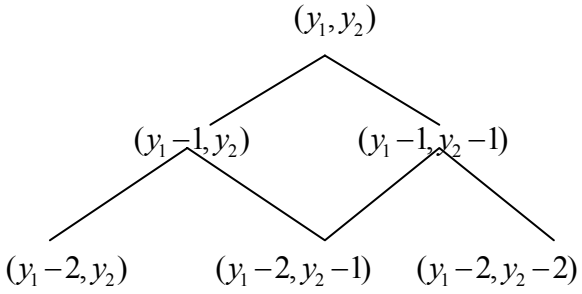


Figure 5.3: Flat algorithm (stage 2)

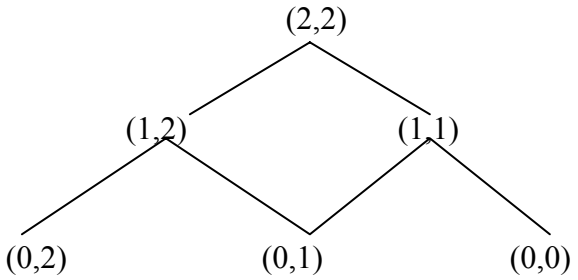


Figure 5.4: Calculating $p(2,2)$ using the Flat algorithm

Detail description of the Flat algorithm for n-variate common covariance structure can be found in Tsiamyrtzis and Karlis (2004). We wrote SPLUS/R functions to calculate the probability function of a trivariate Poisson distribution and a bivariate Poisson distribution using the Flat algorithm and the recurrence relationships.

5.3 Multivariate Poisson finite mixture models

The main idea (Vermunt et al., 2002 and McLachlan et al., 1988) in model-based clustering, also known as latent class clustering or finite mixture models, is that the observations (in our case weed counts) are assumed to be coming from a mixture of density distributions for which the parameters of the distribution and the mixing proportions and the number of the components are unknown. Therefore, the objective of model-based clustering is to unmix the distributions and to find the most favorable parameters of the distributions, and the number and the mixing proportions of the components, given the underlying data (Fraley and Raftery, 1998).

The history of finite mixture models dates back to the late 19^{th} century (Pearson, 1894). With the arrival of high-speed computers, the finite mixture models inventions began, turning the attention to likelihood estimation of the parameters in a mixture distribution (McLachlan et al., 1988). In particular, the explanation of the EM algorithm (expectation_maximization) by Dempster et al., (1977) has given a new motivation to the research of finite mixture models. Since then, a wide range of literature has been published on this topic, even though most of publications date from 1985 and onwards (Brijs, 2002).

Finite mixture models have demonstrated clustering in several practical applications, including character recognition (Murtagh and Raftery, 1984); tissue segmentation (Banfield and Raftery, 1993); minefield and seismic fault detection (Dasgupta and

Raftery, 1998); identification of textile flaws from images (Campbell et al., 1997); classification of astronomical data (Celeux and Govaert, 1995); and classification of radar images (Fjørtoft et al., 2003).

The next section will provide an overview of the general formulation of the finite mixture model and is mainly drawn from books and review articles (McLachlan and Basford, 1988; McLachlan and Peel, 2000; Titterington et al., 1985 and Titterington, 1990).

5.3.1 Description of model-based clustering

In general, in model-based clustering, the observed data are assumed to come from several unknown components (segments, components, latent classes or clusters are synonyms and will sometimes be used interchangeably) that are mixed in unknown proportions. The objective is then to 'unmix' the observations and to estimate the parameters of the underlying density distributions within each component. The idea is the observations belong to the same class are alike with respect to the observed variables in the sense that their observed values are considered as coming from a mixture of the same density distributions, whose parameters are unknown quantities to be estimated (McLachlan and Basford, 1988). The density distribution is used to estimate the probability of the observed values of the component variables, conditional on knowing the mixture component from which those values were drawn.

The population of interest thus consists of k subpopulations and the density (or probability function) of the q-dimensional observation $\mathbf{y}$ from the j^{th} ($j = 1,...,k$) subpopulation is $f(\mathbf{y} \mid \theta_j)$ for some unknown vector of parameters θ_j. The interest lies in finding the values of the non-observable vector $\varphi = (\phi_1, \phi_2, ..., \phi_n)$ which contains the component labels for each observation $(i = 1,....,n)$ and $\phi_i = j$ if the i^{th} observation belongs to the j^{th} subpopulation.

Since the component labels are not observed, the conditional density of the vector $\mathbf{y}$ is a mixture of density of the form

$$f(y_i) = \sum_{j=1}^{k} p_j f(y_i \mid \theta_j), \tag{5.13}$$

where $0 < p_j < 1$, $\sum_{j=1}^{k} p_j = 1$ and p_j are the mixing proportions. Note that the mixing proportion is the probability that a randomly selected observation belongs to the j-th component. This is the classical mixture model (McLachlan and Peel, 2000). The purpose of model-based clustering is to estimate the parameters $(p_1,..., p_{k-1}, \theta_1,..., \theta_k)$. The maximum likelihood (ML) estimation approach, estimates the parameters maximizing the loglikelihood:

$$L(y; \theta, p) = \sum_{i=1}^{n} \ln\left(\sum_{j=1}^{k} p_j f(y_i \mid \theta_j) \right). \tag{5.14}$$

But this is not easy since there is often not a closed-form solution for calculating these parameters. Fortunately, due to the finite mixture representation, an expectation-

maximization (EM) algorithm is applicable (McLachlan and Peel, 2000 and Fraley and Raftery, 1998).

For a multivariate finite mixture model, to avoid the computational difficulties, it is often assumed that the observed variables are mutually independent within components (Vermunt et al., 2002). If there are no restrictions on the dependency of variables, the model with multivariate probability density functions is applicable. Sometimes the model-based clustering problem involves estimating a separate set of means, variances, and covariances for each mixture component, which quickly becomes computationally burdensome (Brijs, 2002).

Several types of restrictions can be imposed on the variance-covariance matrix to create the models in between the local independence model and the full covariance model. In some situations, this may be necessary for practical reasons since the unrestricted model may be inadequate. The reason for this inadequacy is that the number of free parameters in the variance-covariance matrix for the full covariance model increases rapidly with the number of mixture components and the number of indicator variables. Therefore, more restricted models are defined by assuming certain pairs of y's to be mutually independent within mixture components by fixing some but not all covariances to zero (Karlis, 2003; Li et al., 1999).

5.3.2 Model-based cluster estimation

The purpose of model-based clustering, described in previous section, is to estimate the parameter vector $\mathbf{\Phi}$. The maximum likelihood (ML) and the maximum a posterior (MAP) estimation (Vermunt et al., 2002) are the two popular methods to estimate this parameter vector. Of these two, maximum likelihood estimation is used in this thesis.

5.3.3 ML estimation with the EM algorithm

One purpose of model-based clustering approach is to estimate the parameters $(p_1,..., p_{k-1}, \theta_1,..., \theta_k)$. Following the maximum likelihood (ML) estimation approach, the estimation involves maximizing the loglikelihood (5.14), as stated earlier. In other words, the idea is to find the optimal values for the parameter vector, say $\mathbf{\Phi}_{\text{opt}}$, such that the observations $y_i \quad (i = 1,..., n)$ are more likely came from $f(y_i \mid \mathbf{\Phi}_{\text{opt}})$ than from $f(y_i \mid \mathbf{\Phi})$ for any other value of $\mathbf{\Phi}$ (McLachlan and Peel, 2000)**.**

To maximize this loglikelihood, different approaches such as Newton-Raphson algorithm (McHugh, 1956), expectation-maximization (EM) (Dempster et al., 1977; McLachlan and Krishnan, 1997) algorithm etc. can be used. Most of software either uses Newton-Raphson algorithm or expectation-maximization (EM) algorithm, or a combination of both. Most recent techniques increasing in popularity are the stochastic EM method (Diebolt, 1996) and MCMC (Markov Chain Monte Carlo) (Robert, 1996). Moreover, since the EM is relatively slow, recent research efforts focus on modifying

the EM algorithms for use on very large data sets (McLachlan and Peel, 2000). Newton-Raphson algorithm requires fewer iterations than the EM algorithm (McLachlan and Basford, 1988). Quadratic convergence is regarded as the major strength of the Newton-Raphson method. Furthermore, because of its computational straightforwardness, the EM algorithm is the most extensively used (Titterington, 1990). Later in this chapter, a detailed version of the EM for the multivariate Poisson finite mixture models and the multivariate Poisson hidden Markov model is provided. At this moment, the EM can be described as an iterative algorithm that sequentially improves on the sets of starting values of the parameters, and facilitate simultaneous estimation of all model parameters (Dempster et al., 1977; Hasselblad, 1969). More specifically, the observed data y_i is augmented with the unobserved segment membership of subjects z_{ij}, which greatly simplifies the computation of the likelihood instead of maximizing the likelihood over the entire parameter space. More facts about the EM computation can be found in Dempster et al. (1977) and McLachlan et al. (1988). The estimates of the posterior probability w_{ij}, i.e. the posterior probability for subject i belongs to the component j, can be obtained for each observation vector y_i according to Bayes' rule after the estimation of the optimal value of $\mathbf{\Phi}$. In fact, after estimation the density distribution $f(y_i \mid \theta_j)$ within each mixture component j and the component size p_j of each component such that the posterior probability can be calculated as

$$w_{ij} = \frac{p_j f(y_i \mid \theta_j)}{\sum_{j=1}^{k} p_j f(y_i \mid \theta_j)} \ . \tag{5.15}$$

5.3.3.1 Properties of the EM algorithm

The EM algorithm is certainly one of the most accepted algorithms to estimate finite mixture models due to some of its attractive properties listed below:

- The most important advantage of the EM algorithm is surely its convergence towards the optimum parameter values. This means that, given the recent mixture model parameters, a single EM iteration provides new parameter estimates, which are proven to increase the loglikelihood of the model (Dempster et al., 1977; McLachlan et al., 1997). The convergence of the EM algorithm is proven by Meilijson (1989) and Wu (1983).
- The EM algorithm ensured that the estimated parameters are within the required range (admissible range). This means that, for example for the Poisson distribution, the parameter values are zero or positive and cannot take negative values.
- The EM algorithm is fairly easy to program.

However, apart from these appealing properties of the EM algorithm, some limitations have been identified as well:

- The setback with the EM estimation is that the procedure may converge to a local but not a global optimum (McLachlan et al., 1988; Titterington et al., 1985). It is generally accepted that the best way to avoid a local solution is to use multiple sets of starting values for the EM algorithm and to observe the evolution of final likelihood for the different restarts of the EM algorithm.

Another alternative is to use the partitioning of a *k*-means clustering as the initial starting values (McLachlan et al., 1988).

- The EM algorithm generally converges very slowly when compared to other iterative algorithm, such as Newton-Raphson algorithm. The EM converges linearly towards the optimum, while Newton-Raphson converges with quadratic speed towards optimum (Aitkin and Aitkin, 1996).
- Non-convergence to global optimum sometimes is another problem of the EM algorithm. Convergence and the properties of convergence depend heavily on the starting values.
- An important problem, but somewhat ignored in the literature, is the stopping rule for the number of iterations. In fact, the EM is rather sensitive in the sense that different stopping rules can lead to different estimates (Seidel et al., 2000). According to Karlis and Xekalaki (1998), this is caused because at every iteration, the loglikelihood increases by a very small amount and at the same time the estimates can change a lot.
- Even though the EM is more popular, its general principles are well understood and extensively used algorithm, in every problem one has to build the algorithm in a different way.

5.3.4 Determining the number of components or states

In some applications of model-based clustering, there is enough information about the number of components *k* in the mixture model to be specified with sufficient certainty.

For example, where the components correspond to externally existing groups is such situation. Though, often the number of components has to be determined from the data, along with the parameters in the component densities (McLachlan et al., 2000). Regrettably, this crucial problem of finding the optimal number of components in a mixture model has not yet been completely solved (Mackay, 2002). However, a more suitable viewpoint to determine the number of components is based on the use of so called information criteria. The most well-known examples include the AIC (Akaike information criterion) (Akaike, 1974) and the BIC (Bayesian information criterion or Schwarz information criterion) (Schwarz, 1978). The formulas are:

$$AIC = L_k - d_k$$

$$BIC = L_k - \ln(n)\frac{d_k}{2},$$

where L_k - the value of maximized loglikelihood for a model with k components and d_k - the number of free parameters in the model with k components and n is the number of observations.

Information criteria are goodness of fit measures, which consider model parsimony. The main idea is that the increase of the loglikelihood of the mixture model L_k on a particular dataset of size n, penalized by the increased number of parameters d_k needed to create this increase of fit. A larger criterion indicates a better model in comparison with another. In spite of this, it should be noted that several other criteria exists. AIC and BIC criterions have been used to determine the number of states in a hidden Markov model (Leroux and Puterman, 1992).

5.3.5 Estimation for the multivariate Poisson finite mixture models

5.3.5.1 The EM algorithm

The EM algorithm is a popular algorithm for the ML estimation in statistical applications (Dempster et al., 1977; McLachlan and Krishnan, 1997). It is appropriate to the problems with missing values or problems that can be seen as containing missing values. Suppose that there are observed data Y_{obs} and unobservable/missing data Y_{mis}, which are perhaps missing values or even non-observable latent variables. The idea is to augment the observed and the unobserved data, taking the complete data $Y_{com} = (Y_{obs}, Y_{mis})$. The key idea of this algorithm is to iterate between two steps. The first step, the E-step, computes the conditional expectation of the complete data loglikelihood with respect to the missing data, while the second step, the M-step, maximizes the complete data likelihood.

Consider the multivariate reduction proposed earlier (section 5.1) in this thesis. The observed data are the q-dimensional vectors $Y_i = (Y_{1i}, Y_{2i}, Y_{3i})$. The standard data argumentation is used for finite mixture models by introducing as latent variables the vectors $Z_i = (Z_{i1}, Z_{i2}, ..., Z_{ik})$ that correspond to the component memberships with $Z_{ij} = 1$ if the i-th observation belongs to the j^{th} component, and 0 otherwise. Furthermore, some more latent variables are introduced as follows: The component specific latent variables, i.e. for the j^{th} component are introduced using the unobservable vectors $X_i^j = (X_{1i}^j, X_{2i}^j, X_{3i}^j, X_{12i}^j, X_{13i}^j, X_{23i}^j)$, where the superscript

indicates the component, and the variables are the latent variables used to construct the model in section 5.1. Thus, the complete data are the vectors (Y_i, X_i, Z_i). The vector of parameters is defined by $\mathbf{\Phi}$, and then the complete loglikelihood takes the following form:

$$L(\mathbf{\Phi}) = \sum_{i=1}^{n}\sum_{j=1}^{k} Z_{ij}(\log p_j + \log \prod_{t\in\Omega} f(X_{ti}^j \mid \theta_{tj}))$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{k} Z_{ij} \log p_j + \sum_{i=1}^{n}\sum_{j=1}^{k} Z_{ij} \sum_{t\in\Omega} (-\theta_{tj} + X_{ti}^j \log \theta_{tj} - \log X_{ti}^j !), \qquad (5.16)$$

where $\Omega = \{1,2,3,12,13,23\}$. The relevant part of the complete likelihood is give by $\sum_{i=1}^{n}\sum_{j=1}^{k}(-Z_{ij}\theta_{tj} + Z_{ij}X_{ti}^j \log\theta_{tj})$ and hence, one needs the expectations $E(Z_{ij})$ and $E(X_{ti}^j Z_{ij})$. However for the latter, since Z_{ij} is a binary random variable, when X_{ti}^j is 0 if the observation does not belong to the j^{th} component and takes the value X_{ti}^j if Z_{ij}=1. Thus $E[X_{ti}^j Z_{ij}] = p(Z_{ij})E[X_{ti}^j \mid Z_{ij} = 1]$.

The $E[X_{ti}^j \mid Z_{ij} = 1]$ is the expectation of the latent variable X_{ti}^j given that it belongs to the j^{th} component. Thus, at the E-step one needs the expectations $E[Z_{ij} \mid Y_i, \mathbf{\Phi}]$ for $i = 1,...,n$, $j = 1,...,k$ and $E[X_{ti}^j \mid Y_i, Z_{ij} = 1, \mathbf{\Phi}]$ for $i = 1,...,n$, $j = 1,...,k$ and $t \in \Omega$.

More formally, the procedure can be described as follows:

E-step: Using the current values of the parameters calculate

$$w_{ij} = E[Z_{ij} \mid Y_i, \mathbf{\Phi}] = p_j \frac{p(y_i \mid \theta_j)}{p(y_i)}, \quad i = 1,...,n, \quad j = 1,...,k. \qquad (5.17)$$

$$E[X_{12i}^j \mid Y_i, Z_{ij} = 1, \boldsymbol{\Phi}] = d_{12i}^j$$

$$= \sum_{r=0}^{\min(y_{1i}, y_{2i})} r \, \mathrm{P}[X_{12i}^j = r \mid y_{1i}, y_{2i}, Z_{ij} = 1, \boldsymbol{\Phi}]$$

$$= \sum_{r=0}^{\min(y_{1i}, y_{2i})} \frac{r \, \mathrm{P}[X_{12i}^j = r, y_{1i}, y_{2i} \mid Z_{ij} = 1, \boldsymbol{\Phi}]}{\mathrm{P}[y_{1i}, y_{2i} \mid Z_{ij} = 1, \boldsymbol{\Phi}]}$$

$$= \sum_{r=0}^{\min(y_{1i}, y_{2i})} \frac{r Po(y_{1i} - r \mid \theta_{1j}) Po(y_{2i} - r \mid \theta_{2j}) Po(r \mid \theta_{12j})}{p(y_i \mid \theta_j)}. \tag{5.18}$$

The corresponding expressions for $E[X_{13i}^j \mid Y_i, Z_{ij} = 1, \boldsymbol{\Phi}] = d_{13i}^j$ and

$E[X_{23i}^j \mid Y_i, Z_{ij} = 1, \boldsymbol{\Phi}] = d_{23i}^j$ follow by analogy. Then

$$E[X_{1i}^j \mid Y_i, Z_{ij} = 1, \boldsymbol{\Phi}] = d_{1i}^j = y_{1i} - d_{12i}^j - d_{13i}^j$$

$$E[X_{2i}^j \mid Y_i, Z_{ij} = 1, \boldsymbol{\Phi}] = d_{2i}^j = y_{2i} - d_{12i}^j - d_{23i}^j$$

$$E[X_{3i}^j \mid Y_i, Z_{ij} = 1, \boldsymbol{\Phi}] = d_{3i}^j = y_{3i} - d_{13i}^j - d_{23i}^j.$$

M-step: Update the parameters

$$p_j = \frac{\sum_{i=1}^n w_{ij}}{n} \text{ and } \theta_{tj} = \frac{\sum_{i=1}^n w_{ij} d_{ti}^j}{\sum_{i=1}^n w_{ij}}, \text{ for } j = 1, \ldots, k, \; t \in \Omega. \tag{5.19}$$

If some convergence criterion is satisfied, stop iterating; otherwise go back to the E-step. Here the following stopping criterion is used. $\left| \frac{L(k+1) - L(k)}{L(k)} \right| < 10^{-12}$, where $L(k)$ is the loglikelihood at the k^{th} iteration. The similarities with the standard EM algorithm for the finite mixture are straightforward. The quantities w_{ij} at the termination of the algorithm are the posterior probabilities that the i^{th} observation

belongs to the j^{th} cluster, and thus, they can be used to assign the observations to the cluster with higher posterior probability.

This clustering model is also suitable for databases with large amounts of records (Brijs et al., 2004). In fact, even with a very large database, the clustering is done without any additional effort. In general, to examine the suitability of this algorithm, two issues should be taken into account. These issues are the dimensions of the problem and the covariance structure. In fact, it is well known that the speed of the EM algorithm depends on the 'missing' information. One could measure the missing information as the ratio of the observed information to the missing information, which is related to the number of latent variables introduced. Adding more latent variables leads to more 'missing' information and thus adds more computing time.

The above fact is true as far as the number of dimensions is concerned (Brijs et al., 2004). More dimensions lead to more latent variables. If the structure is not complicated, the algorithm will perform relatively the same, but if the structure is more complicated, then more computational effort is needed. In this thesis, the EM algorithm is fully described for the case of two-way interactions (section 5.3.5.1).

5.4 Multivariate Poisson hidden Markov models

Hidden Markov models (or Markov dependent finite mixture models) take a broad view of mixture distributions by introducing serial correlation through the sequence of unseen parameter values $\boldsymbol{\lambda}_j$. In particular, this sequence is assumed to follow a Markov chain with stationary transition probabilities. Formally, let $\{S_i\}$ be a Markov chain with states denoted $1,...,m$ and stationary transition probabilities. Then $\mathbf{y}_i$ are assumed to be conditionally independent given S_i, with conditional densities $f(\mathbf{y}_i;\boldsymbol{\lambda}_{S_i})$. To fit such a model, the transition probabilities must be estimated along with the component parameters $\boldsymbol{\lambda}_j$. Details about the univariate hidden Markov model (or Markov dependent finite mixture models) were described in Chapter 2 and 3.

5.4.1 Notations and description of multivariate setting

The following notations are used through out this section and do not refer to the notations in other sections.

y_{ij} = Measurement of the i^{th} variable on the j^{th} item.

$$\mathbf{Y} = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1j} & \cdots & y_{1n} \\ y_{21} & y_{22} & \cdots & y_{2j} & \cdots & y_{2n} \\ y_{31} & y_{32} & \cdots & y_{3j} & \cdots & y_{3n} \end{bmatrix}.$$

The observation sequences of three variables are denoted by $\mathbf{Y}_1, \mathbf{Y}_2, \mathbf{Y}_3$ row vectors (with uppercase letters)

$$\mathbf{Y}_1 = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1j} & \cdots & y_{1n} \end{bmatrix}.$$

$$\mathbf{Y}_2 = \begin{bmatrix} y_{21} & y_{22} & \cdots & y_{2j} & \cdots & y_{2n} \end{bmatrix}.$$

$$\mathbf{Y}_3 = \begin{bmatrix} y_{31} & y_{32} & \cdots & y_{3j} & \cdots & y_{3n} \end{bmatrix}.$$

To denote the observation sets we will use column vectors $\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_n$ (with lowercase letters)

$$\mathbf{y}_1 = \begin{bmatrix} y_{11} \\ y_{21} \\ y_{31} \end{bmatrix} \quad \mathbf{y}_2 = \begin{bmatrix} y_{12} \\ y_{22} \\ y_{32} \end{bmatrix} \quad \cdots \quad \mathbf{y}_n = \begin{bmatrix} y_{1n} \\ y_{2n} \\ y_{3n} \end{bmatrix}.$$

Then $\mathbf{Y} = \begin{bmatrix} \mathbf{y}_1 & \mathbf{y}_2 & \cdots & \mathbf{y}_j & \cdots & \mathbf{y}_n \end{bmatrix}$ where $\mathbf{y}_n$ is a trivariate observation.

5.4.2 Estimation for the multivariate Poisson hidden Markov models (Extension of the univariate Markov-dependent mixture model by Leroux and Puterman, 1992)

Let $\mathbf{Y} = \begin{bmatrix} \mathbf{y}_1 & \mathbf{y}_2 & . & . & . & \mathbf{y}_j & . & . & . & \mathbf{y}_n \end{bmatrix}$ be the realization of a hidden Markov model with original *m* state Markov Chain $\{S_i\}$. Define $\mathbf{\Phi}$ by $(P_{11}, P_{12}, ..., P_{mm}, \boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, ..., \boldsymbol{\lambda}_m)$

where $P_{jk} = \Pr(S_i = k \mid S_{i-1} = j), \quad 1 \le j,k \le m$ denote the stationary transition probabilities of $\{S_i\}$ and $\boldsymbol{\lambda}_j = \begin{bmatrix} \lambda_{11i} & \lambda_{12i} & \lambda_{13i} & \lambda_{22i} & \lambda_{23i} & \lambda_{33i} \end{bmatrix}$, where $j = 1,...,m$. The likelihood for $\boldsymbol{\Phi}$ is

$$L(\boldsymbol{\Phi} \mid \mathbf{y}_1, \mathbf{y}_{2,}....., \mathbf{y}_n) = \sum_{S_1=1}^{m} ... \sum_{S_n=1}^{m} P_{s_1}^{(1)} f(\mathbf{y}_1; \boldsymbol{\lambda}_{S_1}(\boldsymbol{\Phi})) \prod_{i=2}^{n} P_{S_{i-1}S_i}(\boldsymbol{\Phi}) f(\mathbf{y}_i; \boldsymbol{\lambda}_{S_i}(\boldsymbol{\Phi})), \tag{5.20}$$

where $P_j^{(1)} = \Pr(S_1 = j)$ denote the initial probabilities of $\{S_i\}$. Leroux and Puterman (1992) discussed in their paper that $L(\boldsymbol{\Phi} \mid \mathbf{y}_1, \mathbf{y}_{2,}....., \mathbf{y}_n)$ is a convex mixture of likelihood values obtained with a fixed initial state (i.e. with $P_j^{(1)} = 1$ for some *j*), concurrently maximization of $L(\boldsymbol{\Phi} \mid \mathbf{y}_1, \mathbf{y}_{2,}....., \mathbf{y}_n)$ over $\boldsymbol{\Phi}$ and $(P_1^{(1)},...,P_m^{(1)})$ can be accomplished by maximization over $\boldsymbol{\Phi}$ with a fixed initial state. Thus, it follows that the $P_j^{(1)}$ are known. Cappé (2001) explained that with a single training sequence, the initial distribution is a parameter that has not much effect and the initial distribution cannot be estimates consistently. Taking the above reason into account, it is assumed that the initial distribution is uniform (equal probabilities for all states of the model). In this thesis, initial Uniform distribution is assumed.

5.4.2.1 The EM algorithm

The EM algorithm can be applied to determine the likelihood maximization for the multivariate Poisson hidden Markov model, almost as simply as for the independent

mixture model. The loglikelihood function for $(s_i, \mathbf{y}_i)$, $i = 1,...,n$ (called the complete-data loglikelihood) is $\log L^c(\mathbf{\Phi} \mid s_1, \mathbf{y}_1, ..., s_n, \mathbf{y}_n)$

$$= \log P^{(1)}_{s_1} + \sum_{i=2}^{n}\sum_{j=1}^{m}\sum_{k=1}^{m} v_{jk}(i) \log P_{jk} + \sum_{i=1}^{n}\sum_{j=1}^{m} u_j(i) \log f(\mathbf{y}_i; \boldsymbol{\lambda}_j), \tag{5.21}$$

where $u_j(i) = 1$ if $S_i = j$ and 0 otherwise, and

$v_{jk}(i) = 1$, if a transition from j to k occurred at i (i.e; $S_{i-1} = j, S_i = k$) and 0 otherwise. ($\mathbf{\Phi}$ is suppressed for simplicity of notation).

This loglikelihood function consists of two parts, the loglikelihood for a Markov chain, depending merely on the transition probabilities P_{jk}, and the loglikelihood for independent observations, depending only on the parameters $\boldsymbol{\lambda}_1, ..., \boldsymbol{\lambda}_m$. Note that when P_{jk} is independent of j, (5.21) gives the complete-data likelihood for the independent case, so that the independent model is nested in the hidden Markov model.

The M-step requires maximization of $E[\log L^c(\mathbf{\Phi}) \mid \mathbf{y}_1, ..., \mathbf{y}_n]$, which is obtained by replacing the components of the missing data by their conditional means.

$$\hat{v}_{jk}(i) = E[v_{jk}(i) \mid \mathbf{y}_1, ..., \mathbf{y}_n] = \mathrm{P}[S_{i-1} = j, S_i = k \mid \mathbf{y}_1, ..., \mathbf{y}_n] \tag{5.22}$$

and

$$\hat{u}_j(i) = E[u_j(i) \mid \mathbf{y}_1, ..., \mathbf{y}_n] = \mathrm{P}[S_i = j \mid \mathbf{y}_1, ..., \mathbf{y}_n]. \tag{5.23}$$

The transition probabilities are obtained using following formula:

$$P_{jk} = \frac{\sum_{i=2}^{n} \hat{v}_{jk}(i)}{\sum_{i=2}^{n}\sum_{l=1}^{m} \hat{v}_{jl}(i)} \quad . \tag{5.24}$$

These equations, similar to the equations 5.17 for the mixing proportions in a mixture distribution, can be thought of as weighted empirical relative frequencies. The maximizing values of $\boldsymbol{\lambda}_j$ are obtained exactly as for independent observations. The algorithm is terminated when the changes in parameter estimates are small.

5.4.2.2 The forward-backward algorithm

The forward-backward algorithm is again an extension of univariate case (Chapter 2 and 3) to a multivariate case. The forward-backward algorithm is used to calculate the conditional probabilities $\hat{u}_j(i)$ and $\hat{v}_{jk}(i)$. It is based on simple recursive formulae for the forward variable

$\alpha_j(i) = P[\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_n, S_i = j]$ and the backward variable

$$\beta_j(i) = P[\mathbf{y}_{i+1}, ..., \mathbf{y}_n \mid S_i = j] \tag{5.25}$$

which yield the quantities of interest by

$$\hat{u}_j(i) = \frac{\alpha_j(i)\beta_j(i)}{\sum_l \alpha_l(n)} = \frac{\alpha_j(i)\beta_j(i)}{\sum_{j=1}^{m} \alpha_j(i)\beta_j(i)} \quad \text{and}$$

$$\hat{v}_{jk}(i) = \frac{P_{jk} f(\mathbf{y}_i; \boldsymbol{\lambda}_k)\alpha_j(i-1)\beta_k(i)}{\sum_l \alpha_l(n)} \quad . \tag{5.26}$$

The $\alpha_j(i)$ and $\beta_j(i)$ are calculated recursively in *i* using following formulae:

$$\alpha_j(i) = \sum_{k=1}^{m} \alpha_k(i-1) P_{kj} f(\mathbf{y}_i; \boldsymbol{\lambda}_j) \tag{5.27}$$

$[\alpha_j(1) = P_j^{(1)} f(\mathbf{y}_1; \boldsymbol{\lambda}_j), j = 1,...,m]$, and

$$\beta_j(i) = \sum_{k=1}^{m} P_{jk} f(\mathbf{y}_{i+1}; \boldsymbol{\lambda}_k) \beta_k(i+1) \tag{5.28}$$

$$[\beta_j(n) = 1, j = 1,...,m].$$

Note that the α's are computed by a forward pass through the observations and the β's by a backward pass after evaluating the α's. The likelihood is then simply calculated by the expression $\sum_{j=1}^{m} \alpha_j(n)$.

The calculations of $X_1, X_2, X_3, X_{12}, X_{13}$, and X_{23} can be carried out using the same formulas explained in section 5.2.2. The multivariate Poisson model is defined as

$$\begin{aligned} \mathbf{Y_1} &= \mathbf{X_1} + \mathbf{X_{12}} + \mathbf{X_{13}} \\ \mathbf{Y_2} &= \mathbf{X_2} + \mathbf{X_{12}} + \mathbf{X_{23}} \\ \mathbf{Y_3} &= \mathbf{X_3} + \mathbf{X_{13}} + \mathbf{X_{23}}. \end{aligned} \tag{5.29}$$

E-step: Using the current values of the parameters calculate

$$E[\mathbf{X}_{12i}^j \mid \mathbf{Y}, u_j(i) = 1, \boldsymbol{\Phi}] = d_{12i}^j$$

$$= \sum_{r=0}^{\min(y_{1i}, y_{2i})} \frac{r Po(y_{1i} - r \mid \lambda_{1j}) Po(y_{2i} - r \mid \lambda_{2j}) Po(r \mid \lambda_{12j})}{f(\mathbf{y}_i \mid \boldsymbol{\lambda}_j)}. \tag{5.30}$$

The corresponding expressions for $E[\mathbf{X}_{13i}^j \mid \mathbf{Y}, u_j(i) = 1, \boldsymbol{\Phi}] = d_{13i}^j$ and $E[\mathbf{X}_{23i}^j \mid \mathbf{Y}, u_j(i) = 1, \boldsymbol{\Phi}] = d_{23i}^j$ follow by analogy. Then

$$E[\mathbf{X}_{1i}^j \mid \mathbf{Y}, u_j(i)=1, \mathbf{\Phi}] = d_{1i}^j = y_{1i} - d_{12i}^j - d_{13i}^j$$

$$E[\mathbf{X}_{2i}^j \mid \mathbf{Y}, u_j(i)=1, \mathbf{\Phi}] = d_{2i}^j = y_{2i} - d_{12i}^j - d_{23i}^j$$

$$E[\mathbf{X}_{3i}^j \mid \mathbf{Y}, u_j(i)=1, \mathbf{\Phi}] = d_{3i}^j = y_{3i} - d_{13i}^j - d_{23i}^j .$$

Let denote $\mathbf{d}_i^j = (d_{1i}^j, d_{2i}^j, d_{3i}^j, d_{12i}^j, d_{13i}^j, d_{23i}^j),\ i=1,...,n,\ j=1,...,m.$

Then M-step computes the posteriori probabilities using the following equation.

$$\widehat{u}_j(i) = P[S_i = j \mid \mathbf{Y} = \mathbf{y}] = \frac{\alpha_j(i)\beta_j(i)}{\sum_{l=1}^{m}\alpha_l(n)} = \frac{\alpha_j(i)\beta_j(i)}{\sum_{j=1}^{m}\alpha_j(i)\beta_j(i)} \tag{5.31}$$

and then re-estimate the rates as follows:

$$\hat{\boldsymbol{\lambda}}_j = \frac{\sum_{i=1}^{n}\widehat{u}_j(i)\mathbf{d}_i^j}{\sum_{i=1}^{n}\widehat{u}_j(i)}, \qquad j=1,...,m. \tag{5.32}$$

The M-step will give the parameter estimates $\boldsymbol{\lambda}_1,...,\boldsymbol{\lambda}_m$ for the k^{th} iteration and then go back, and repeat the algorithm until the convergence criterion is met.

We extended the univariate Markov-dependent Poisson mixture model to a multivariate Poisson model (bivariate and trivariate). We carried out Splus/R codes for the analysis of the multivariate Poisson hidden Markov model according to sections 5.4.2, 5.4.2.1 and 5.4.2.2.

5.5 Bootstrap approach to standard error approximation

The standard error of the parameter estimates in a mixture model can be obtained by approximating the covariance matrix of $\hat{\boldsymbol{\Phi}}$ using the inverse of the information matrix. It is important to mention that these estimates of the covariance matrix of the maximum likelihood estimation based on the expected or observed information matrices are guaranteed to be valid inferentially only asymptotically (McLachlan and Peel, 2000). In particular, for mixture models, it is recognized that the sample size n has to be very large before the asymptotic theory of maximum likelihood applies (McLachlan and Peel, 2000). Since our sample size is not too large, we can use a resampling approach to this problem, the bootstrap method. Basford et al., (1997) and Peel (1998) compared the bootstrap and information-based approaches for some normal mixture models. They found that unless the sample size was very large, the standard errors found by the information-based approach were too unstable to be recommended. In such situations the bootstrap approach is recommended and we use this approach in this thesis.

The bootstrap approach of calculating the standard error is explained by McLachlan and Peel (2000). Tthe bootstrap method was first introduced by Efron (1979). Thereafter the series of articles and books by Efron (1982), Efron and Tibshirani (1993), Davison and Hinkley (1997), Chernick (1999) were published. Over the past twenty-five years, the bootstrap method has become one of the most admired developments in statistics.

The bootstrap method is a powerful and important technique permits the variability in a random quantity to be assessed using just the data at hand (McLachlan and Peel, 2000). An estimate $\hat{F}$ of the underlying distribution is formed from the observed data $\mathbf{Y}$. Conditional on the latter, the sampling distribution of the random quantity of interest with F replaced by $\hat{F}$ defines its so-called bootstrap distribution, which provides an approximation to its true distribution. It is assumed that $\hat{F}$ has been so formed that the stochastic structure of the model has been preserved. Usually, it is impossible to express the bootstrap distribution in a simple form, and it must be approximated by Monte Carlo methods whereby pseudo-random samples (bootstrap samples) are drawn from $\hat{F}$. If a parametric form is adopted for the distribution function of $\mathbf{Y}$, where $\mathbf{\Phi}$ denotes the vector of unknown parameters, and then the parametric bootstrap uses an estimate $\hat{\mathbf{\Phi}}$ formed from $\mathbf{y}$ in place of $\mathbf{\Phi}$. That is, if we write F as F_{Φ} to signify its dependence on $\mathbf{\Phi}$, then the bootstrap data are generated from $\hat{F} = F_{\hat{\Phi}}$.

McLachlan and Peel (2000) explained that the standard error estimation of $\hat{\mathbf{\Phi}}$ could be stated using the bootstrap method by the following steps:

Step 1: The new data, $\mathbf{Y}^*$, called the bootstrap sample, is generated according to $\hat{F}$, an estimate of the distribution formed from the original observed data $\mathbf{Y}$. That is, in the case where $\mathbf{Y}$ contains the observed values of a random sample of size *n*, $\mathbf{Y}^*$ consists of the observed values of the random sample

$$\mathbf{Y_1^*}, \mathbf{Y_2^*}, ..., \mathbf{Y_n^*} \overset{i.i.d.}{\sim} \hat{F},$$

where the estimate $\hat{F}$ (now denoting the distribution function of a single observation $\mathbf{Y}_j$) is held fixed at its observed value.

Step 2: The EM algorithm is applied to the bootstrap observed data $\mathbf{Y}^*$ to compute the maximum likelihood estimates for this dataset, $\hat{\mathbf{\Phi}}^*$.

Step 3: The bootstrap covariance matrix of $\hat{\mathbf{\Phi}}^*$ is given by

$$\text{cov}^*(\hat{\mathbf{\Phi}}^*) = E^*[\{\hat{\mathbf{\Phi}} - E^*(\hat{\mathbf{\Phi}})\}\{\hat{\mathbf{\Phi}}^* - E^*(\hat{\mathbf{\Phi}}^*)\}^T], \tag{5.33}$$

where E^* denotes expectation over the bootstrap distribution specified by $\hat{F}$.

The bootstrap covariance matrix can be approximated by Monte Carlo methods. Steps (1) and (2) are repeated independently several times (say, B) to give B independent realizations of $\hat{\mathbf{\Phi}}^*$, denoted by $\hat{\mathbf{\Phi}}_1^*,....,\hat{\mathbf{\Phi}}_B^*$. Then (5.33) can be approximated by the sample covariance matrix of these B bootstrap replications to give

$$\text{cov}^*(\hat{\mathbf{\Phi}}^*) \approx \frac{\sum_{b=1}^{B}(\hat{\mathbf{\Phi}}_b^* - \overline{\hat{\mathbf{\Phi}}^*})(\hat{\mathbf{\Phi}}_b^* - \overline{\hat{\mathbf{\Phi}}^*})^T}{(B-1)} \tag{5.34}$$

where $\overline{\hat{\mathbf{\Phi}}^*} = \frac{\sum_{b=1}^{B}\hat{\mathbf{\Phi}}^*}{B}$.

The standard error of the i^{th} element of $\hat{\mathbf{\Phi}}$ can be estimated by the positive square toot of the i^{th} diagonal element of (5.34). It has been demonstrated that 50 to 100 bootstrap

replications are generally sufficient for the standard error estimation (Efron and Tibshirani, 1993).

On the identifiability of a mixture model, if the component densities of the mixture belong to the same parametric family, then the likelihood does not change under a permutation of the component labels in the parameter $\boldsymbol{\Phi}$ and hence neither does its maximum likelihood estimate $\widehat{\boldsymbol{\Phi}}$. This raises the question of whether the so-called label-switching problem (for example, what you have as the first cluster now will be the second cluster in the next sample and so on) occurs in the generation of the bootstrap replications of the maximum likelihood estimation, as in Monte Carlo Markov chain computations involving mixture models. McLachlan and Peel (2000) explained that according to their experience it has not arisen, as they always take the maximum likelihood estimate $\widehat{\boldsymbol{\Phi}}$ calculated from the original data to be the initial value of parameter in applying the EM algorithm to each bootstrap sample.

The following steps were used to calculate the bootstrapped standard errors for both models:

(a) Multivariate Poisson finite mixture model and

(b) Multivariate Poisson hidden Markov model.

Step 1: Using estimated means and transition probabilities/or mixing proportions from different states/or components simulate the mixture distribution of data.

Step 2: Then take a bootstrap sample (with replacement) of size equal to the original sample size and estimate the parameters using the EM algorithm.

Step 3: Take at least 100 bootstrap samples and estimate the parameters.

Step 4: Finally using these 100 bootstrap parameters calculates the standard errors of the estimates.

5.6 Splus/R codes for the multivariate Poisson hidden Markov model

We contributed to the hidden Markov model research area by developing Splus/R codes for the analysis of the multivariate Poisson hidden Markov model. Splus/R codes are written to estimate the multivariate Poisson hidden Markov model using the EM algorithm and the forward-backward procedure and the estimation of bootstrapped standard errors. The estimated parameters were used to calculate the goodness of fit measures mention in this thesis: the entropy criterion (section 6.5) and the estimated unconditional variance-covariance matrix (section 7.3). Splus/R programs (see Appendix) of this thesis are available on request from the author.

5.7 Loglinear analysis

Loglinear models were used to identify the covariance structure in this thesis. The loglinear model is a special case of generalized linear model (GLM) for count-type response variables modelled as Poisson data (Agresti, 2002). All generalized linear models have three components. The random component identifies the response variable

Y and assumes a probability distribution for it. The systematic component specifies the explanatory variables used as predictors in the model. The link function describes the functional relationship between the systematic component and the expected value (mean) of the random component. The generalized linear model relates a function of that mean to the explanatory variables through a prediction equation having linear form (Agresti, 2002). More details of the GLM and the loglinear analysis can be found in Agresti (2002).

A generalized linear model using the log link function with a Poisson response is called a loglinear model. The general use is modelling cell counts in contingency tables. The models specify how the expected count depends on levels of the categorical variables for that cell as well as associations and interactions among those variables. To calculate the level of interdependence between two species and higher-order associations, loglinear analysis provides a good statistical background to directly examine the higher-order associations. Loglinear models methodology is mainly applicable when there is no clear distinction between response and explanatory variables, for example, when all the variables are observed simultaneously (Stokes et al., 2000). The loglinear model point of view treats all variables as response variables, and the focus is on statistical independence and dependence. Loglinear modelling of multi-way categorical data is analogous to correlation analysis for normally distributed response variables and is useful in assessing patterns of statistical dependence among the subsets of variables.

The loglinear model is one special case of Generalized Linear Models (GLM) for Poisson distributed data (Agresti, 2002 and Brijs, 2002). Further, loglinear analysis can be considered as an extension of the two-way contingency table to where the conditional relationship between two or more discrete categorical variables is analyzed by taking the natural logarithm of the cell frequencies within the contingency table. Loglinear models are generally used to summarize multi-way contingency tables that involve three or more variables. Therefore, loglinear models are very useful to evaluate the association between variables. PROC CATMOD procedure in SAS software (SAS/STAT, 2003) can be used to fit the models.

The fundamental strategy in loglinear analysis involves fitting models to the observed frequencies in the cross-tabulation of categorical variables. The models can then be represented by a set of expected frequencies that may or may not look like the observed frequencies. Different models can be described in terms of marginal models that they fit and in terms of the constraints they impose on the associations that are present in the data. Using expected frequencies, different models can be fitted and compared that are hierarchical to one another. The idea of modelling is then to choose a preferred model, which is the most suitable model that fits the data. The choice of the preferred model is based on a formal comparison of goodness-of-fit statistics (likelihood ratio test) associated with models that are related hierarchically (i.e. models containing higher order terms also implicitly include all lower order terms).

For the case of two categorical variables, each with two levels (2×2 table with present and absent of the species), to evaluate if an association exists between the variables the following model can be used:

$$\ln(F_{ij}) = \mu + \gamma_i^A + \gamma_j^B + \gamma_{ij}^{AB} \quad . \tag{5.35}$$

$\ln(F_{ij})$ is the log of the expected cell frequency of the cases for cell i, j in the contingency table.

μ is the overall mean of the natural log of the expected frequencies

γ represent the 'effects', which the variables have on the cell frequencies

A and B are two categorical variables

i and j refer to the categories within the variables

Therefore:

γ_i^A = the main effect for variable A

γ_j^B = the main effect for variable B

γ_{ij}^{AB} = the interaction effect for variables A and B.

The model presented by equation (5.35) is called the saturated model. It includes all possible one-way and two-way effects. Given that the saturated model has the same number of effects as there are cells in the contingency table, the expected cell frequencies will always exactly match the observed frequencies, with no degrees of freedom remaining (Agresti, 2002). To find a more parsimonious model that will isolate the effects best explaining the data, a non-saturated model must be discovered. This model could be achieved by setting some of the effect parameters to zero. For instance,

if the effects parameter γ_{ij}^{AB} is set to zero (i.e. assume that variable A has no effect on variable B, or vice versa), the unsaturated model is obtained:

$$\ln(F_{ij}) = \mu + \gamma_i^A + \gamma_j^B . \tag{5.36}$$

Furthermore, it can be said that the models presented above are hierarchically related to each other, i.e. they are nested. In other words, the unsaturated model is nested within the saturated model.

From the collection of unsaturated models that have been fitted, it is required to decide which of the unsaturated models provides the best fit. The likelihood ratio test (G^2) can be carried out to find out the best-fitted model, since the models are nested within each other. If F_{ij} represents the fitted frequency and f_{ij} the observed frequency, then the likelihood ratio test statistic (Agresti, 2002) is denoted by:

$$G^2 = 2\sum_i \sum_j f_{ij} \log\left(\frac{f_{ij}}{F_{ij}}\right). \tag{5.37}$$

The G^2 test is distributed chi-square with degrees of freedom (*df*) equal to the number of cells minus the number of non-redundant parameters (number of model parameters) in the model. In other words, the *df* equals the number of γ parameters set equal to zero. When the models get more complex, the *df* value decreases, with the *df*=0 for the saturated model. As a result, the G^2 tests the residual frequency not accounted for by the effects in the model. (i.e. the γ parameters set equal to zero). Therefore, larger G^2 values indicate that the model does not fit the data well, and thus the model should be

rejected. In this situation, the G^2 test can be used to compare the saturated model with a (smaller) nested model:

$$G^2_{difference} = G^2_{nested} - G^2_{overall} . \quad (5.38)$$

The degrees of freedom (*df*) equal the *df* of the nested model minus the *df* of the saturated model. If the $G^2_{difference}$ is not significant, it means that the more parsimonious nested model is not significantly worse than the saturated model. Then, one should choose the nested model since it is simpler.

This could be easily extended to three variables model each with two levels (with present and absent of the species). The general loglinear model for a three-way table is

$$\ln(F_{ijk}) = \mu + \gamma_i^A + \gamma_j^B + \gamma_k^C + \gamma_{ij}^{AB} + \gamma_{ik}^{AC} + \gamma_{jk}^{BC} + \gamma_{ijk}^{ABC} . \quad (5.39)$$

The total number of non-redundant parameters is the total number of cell counts, which is $2 \times 2 \times 2 = 8$.

CHAPTER 6

RESULTS OF MULTIVARIATE POISSON FINITE MIXTURE MODELS AND MULTIVARIATE POISSON HIDDEN MARKOV MODELS

6.1 Introduction

In this chapter, the results of the multivariate Poisson finite mixture (independent) and the multivariate Poisson hidden Markov models are discussed. The preliminary analysis is presented in sections 6.2 and 6.3. The methodology explained in Chapter 5 is used to calculate the posterior probabilities and to estimate the corresponding parameters. The results of the empirical analysis are presented in section 6.4. A comparison of different model specifications is given in section 6.5.

6.2 Exploratory data analysis

The histogram of species counts for each of the variable and the basic statistics, including the mean and the variance per variable, are illustrated in Figure 6.1 and Table 6.1 respectively. In fact, several important conclusions can be made from these histograms and the basic statistics. First of all, it can be seen from the histograms that the data were discrete integer values (i.e. count data) that can be assumed to model by a Poisson distribution. It is generally accepted in the literature (Johnson et al., 2005) that

the Poisson distribution is well suited to model this kind of data. However, the basic statistics also demonstrate that the data is clearly overdispersed (Table 6.1), i.e. the variance is clearly bigger than the mean and this is a problem when modelling the data with the Poisson distribution. The mean of the Poisson distribution is equal to its variance, can be denoted by single parameter λ, which is not really accurate for the data. The solution to the problem of overdispersion (Leroux and Puterman, 1992) is to assume that the data came from a finite mixture of Poisson distributions, that is, an unknown number of components with different unknown mean species rates.

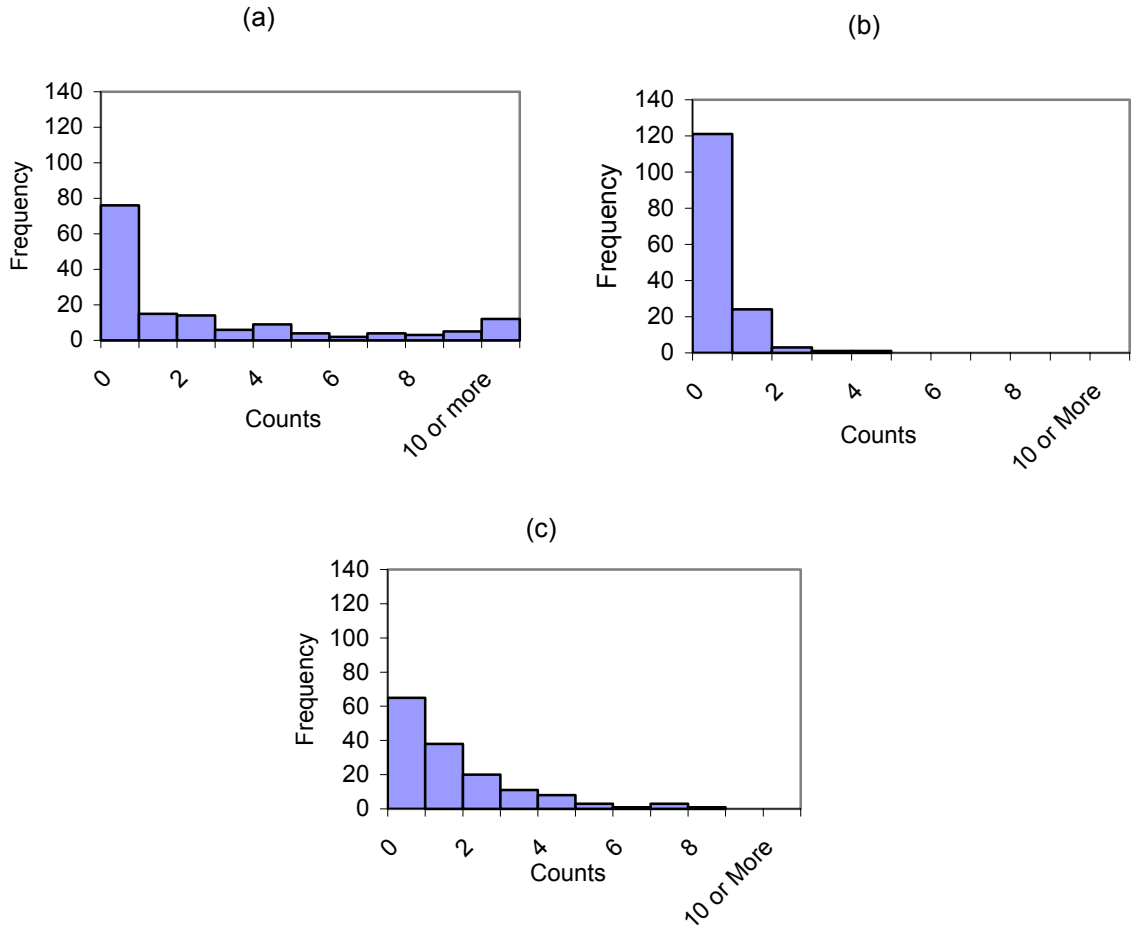


Figure 6.1: Histograms of the species counts: (a) Wild Buckwheat, (b) Dandelion and (c) Wild Oats

Table 6.1: Mean, variance and variance/mean ratio for the three species

Species	Mean	Variance	Variance/Mean
Wild Buckwheat (Species 1097)	1.2867	2.8099	2.1838
Dandelion (Species 1213)	0.2467	0.3481	1.4110
Wild Oats (Species 1509)	2.8200	27.7325	9.8342

Table 6.2: Univariate Poisson mixture models

Number of clusters or states	Univariate Poisson finite mixture models		Univariate Poisson Hidden Markov models	
	AIC	BIC	AIC	BIC
Wild buckwheat (species 1097)	2	2	3	2
Dandelian (species 1213)	1	1	1	1
Wild Oats (species 1509)	5	4	3	3

The univariate analysis was carried out for each species separately to determine how many clusters or states are in each count distribution. The univariate Poisson finite mixture models and univariate hidden Markov models (Leroux and Puterman, 1992) were fitted for each species and AIC and BIC criterions were used to select the number of components of the model. There were different numbers of clusters or states for three species distributions (Table 6.2). The AIC selection was the same compared to the BIC selection method for the most of the models except for two situations. This table gives us an indication that there was more than one cluster or state in species distributions. It is interested to see how many clusters or states were present at the multivariate case.

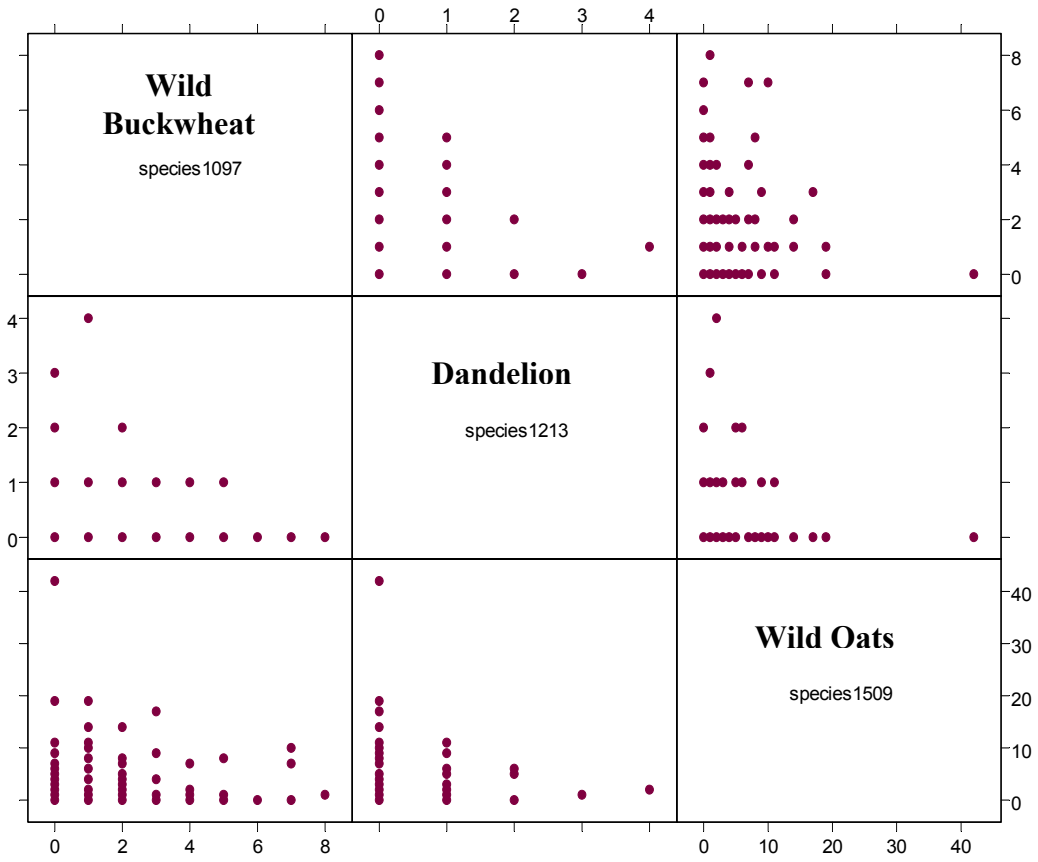


Figure 6.2: Scatter plot matrix for three species

The bivariate correlation analysis (Table 6.3 and Figure 6.2) revealed that there was no statistically significant interaction between variables (all p values >0.05). However, since there may be more complex structure of interactions (i.e., multivariate), the loglinear analysis was carried out on these data to analyze the existence of potentially higher-order interactions between variables.

Table 6.3: Correlation matrix of three species

Correlation	Wild Buckwheat (Species 1097)	Dandelion (Species 1213)	Wild Oats (Species 1509)
Wild Buckwheat (Species 1097)	1	0.0162 (p = 0.844)	0.0119 (p =0.884)
Dandelion (Species 1213)		1	-0.0331 (p = 0.687)
Wild Oats (Species 1509)			1

6.3 Loglinear analysis

The contingency table of the frequency of occurrence (present/absent) of all species combinations of Wild buckwheat, Dandelion and Wild Oats for 150 locations in field #1 is given in Table 6.4. The symbol "0" indicates species was not present at any particular location and the symbol "1" indicates species was present at that location. Table 6.4 illustrates that, out of 150 locations, 34 locations do not contain any of three species, whereas 12 locations contain all of them. Performing a loglinear analysis (SAS/STAT, 2003) on these data described in section 5.7 demonstrates that the saturated model can be significantly reduced to obtain a more suitable, unsaturated model containing less *m*-way interactions. This section has made an attempt to introduce such a model.

Table 6.4: The frequency of occurrence (present/ absent) of the Wild Buckwheat, Dandelion and Wild Oats

Wild buckwheat (Species 1097)	Dandelion (Species 1213)	Wild Oats (Species 1509)	Count
0	0	0	34
1	0	0	28
0	1	0	5
0	0	1	22
1	1	0	7
1	0	1	37
0	1	1	5
1	1	1	12

The likelihood ratio (G^2) test has demonstrated the most suitable model that fits the data only consists of the main effects (Table 6.5).

Table 6.5: The likelihood ratio (G^2) test for the different models of the Wild Buckwheat, Dandelion and Wild Oats counts

Field #1-Effects	G^2	**df**	**P value**
$Y_1+Y_2+Y_3+Y_1Y_2+Y_2Y_3+Y_1Y_3+Y_1Y_2Y_3$	0	0	-
$Y_1+Y_2+Y_3+Y_1Y_2+Y_2Y_3+Y_1Y_3$	0.04	1	0.8414
$Y_1+Y_2+Y_3+Y_1Y_2+Y_2Y_3$	4.23	2	0.1205
$Y_1+Y_2+Y_3+ Y_2Y_3$	5.58	3	0.1341
$Y_1+Y_2+Y_3$	**6.49**	**4**	**0.1654**
$Y_1+ Y_3$	67.13	5	<0.0001

For this three-variate model, the statistical significance of the weed counts interactions between Wild Buckwheat (Y_1), Dandelion (Y_2) and Wild Oats (Y_3) was already studied by means of the loglinear analysis and demonstrated that there were no significant 2-fold interactions. The loglinear analysis is particularly relevant for the development of a simpler multivariate Poisson mixture model for clustering since it helps to discover which interaction terms in the variance/covariance matrix can be set equal to zero. The

p values of the goodness of fit of the models with some two-fold interactions and without any interaction do not differ very much. Therefore, the latent variables $X = (X_1, X_2, X_3, X_{12}, X_{13}, X_{23})$ are decided to keep in the model (i.e. use all two-fold interaction terms). The vector of parameters is now $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3, \theta_{12}, \theta_{13}, \theta_{23})$ and thus the following restricted covariance model can be formulated:

$$
\begin{aligned}
Y_1 &= X_1 + X_{12} + X_{13} \\
Y_2 &= X_2 + X_{12} + X_{23} \\
Y_3 &= X_3 + X_{13} + X_{23}.
\end{aligned}
$$

6.4 Data analysis

In this section, the computational results of the multivariate Poisson finite mixture model and the multivariate Poisson hidden Markov model with the restricted covariance structure is discussed and compared with the results of the local independence model and the common covariance structure. The computational results of the fully saturated multivariate Poisson finite mixture model and the fully saturated multivariate Poisson hidden Markov model will not be discussed since there is no available method to estimate the parameters of the fully saturated multivariate Poisson model in a reliable way. As mentioned in section 5.1, the computation of the fully saturated model involves a great number of summations and parameters to be estimated and this remains a difficulty for calculation of the probability function. As a result, a comparison with the fully saturated covariance model cannot be made.

6.4.1 Results for the different multivariate Poisson finite mixture models

All three models, i.e. the local independence model, the common covariance model, and the model with the restricted covariance structure, were fitted sequentially for 1 to 7 components (k =1,…,7). Furthermore, in order to overcome the famous shortcomings of the EM algorithm, i.e. the dependence on the initial starting values for the model parameters, 10 different sets of starting values were chosen at random. In fact, the mixing proportions (p) were uniform random numbers. These proportions were rescaled so that the summation of all p's is equal to 1. The λ's were generated from a uniform distribution over the range of the data points. For each set of initial values, the algorithm was run for 150 iterations without considering any convergence criterion. Then, the set of initial starting values with the largest loglikelihood was selected. The EM iterations were continued with these selected initial values until the convergence criterion is satisfied, i.e., until the relative change of the loglikelihood between two successive iterations was smaller than 10^{-12}. This procedure is repeated 7 times for each value of k. The number of cluster selection was based on the most well-known information criterion (section 5.3.4), i.e., the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). For the restricted covariance, the independent and the common covariance models d_k is $d_k = 7k-1$, $d_k = 4k-1$ and $d_k = 5k-1$ respectively. The AIC and BIC criterions serve as a guide for the researcher to select the optimal number of components in the data.

Figure 6.3 illustrates the evolution of the loglikelihood, the AIC and the BIC for different components (k =1,…,7) of the local independence multivariate Poisson model. This figure demonstrates that the AIC selects 6 components whereas the BIC selects 5 components. Therefore, in this case, the model with fewer components is selected for interpretation.

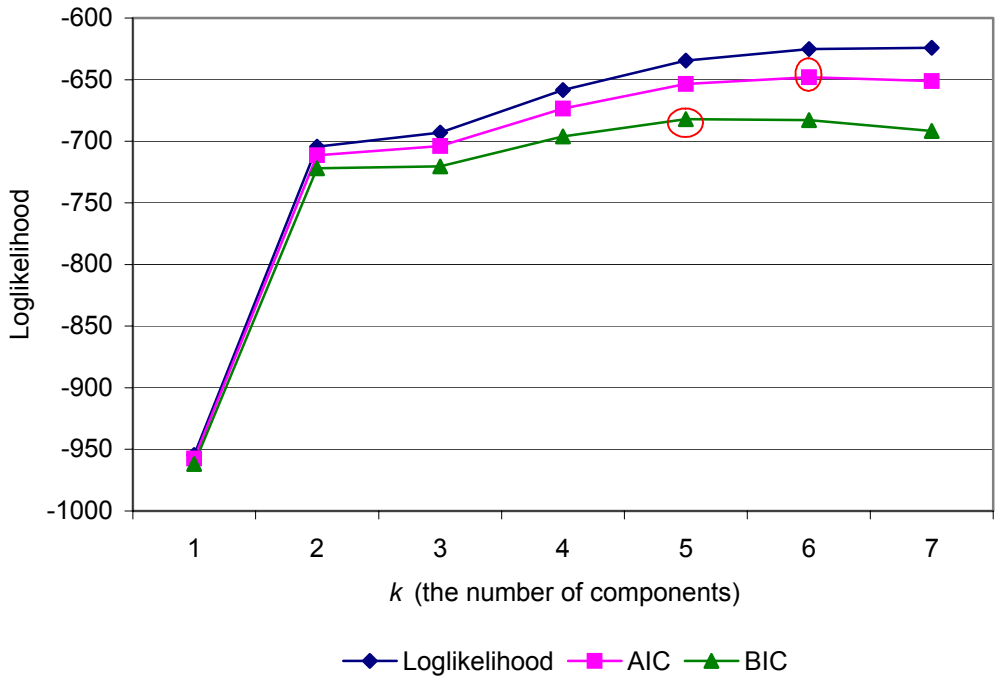


Figure 6.3: Loglikelihood, AIC and BIC against the number of components for the local independence multivariate Poisson finite mixture model

Figure 6.4 illustrates the optimal value of the mixing proportions for the range of models used (values of k from 2 to 7). It can be seen that there is one large component and the rest are small components in all models. In fact, the mixing proportions tend to fluctuate over the different component solutions.

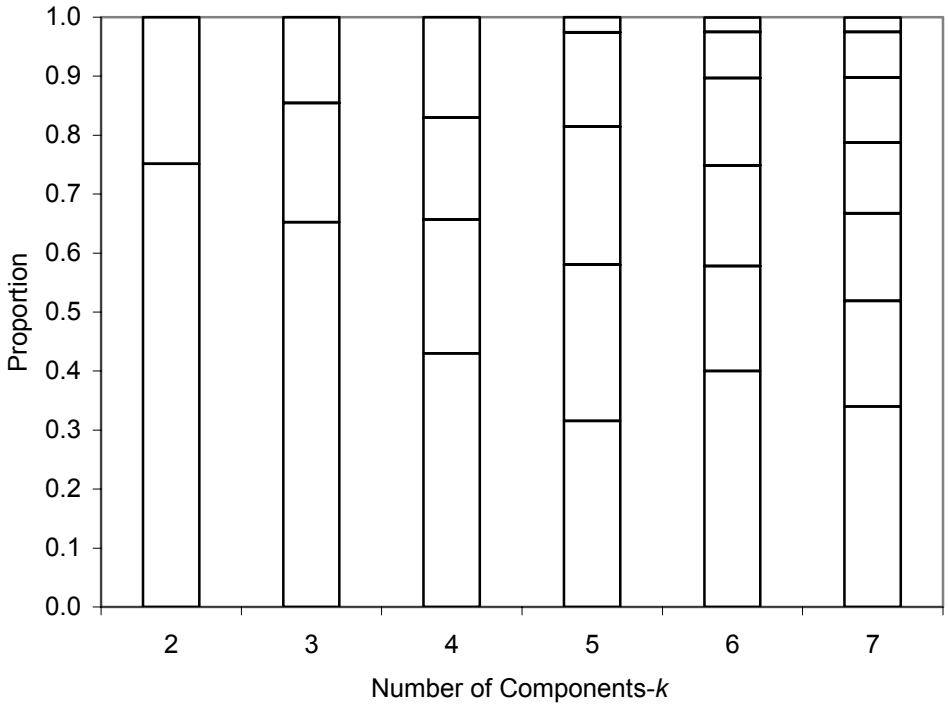


Figure 6.4: The mixing proportions for model solutions with k =2 to 7 components for the local independence multivariate Poisson finite mixture model

Table 6.6 contains the parameter estimates and the bootstrapped standard errors for the model with 5 components. Here the bootstrap standard errors were considered because of small sample size (McLachlan et al., 2000), and therefore, the asymptotic standard errors were not valid. Special care was taken to avoid the label switching (Brijs et al., 2004). This problem can be avoided by adding the relevant constraints, $p_1 \leq p_2 \leq ... \leq p_j$ to the optimization algorithm (p_j's the mixing proportions). For some of the small components with small mixing proportions, the estimated standard errors were large. The parameters with zero estimated values and zero standard errors can be interpreted as zero.

Table 6.6: Parameter estimates (bootstrap standard errors) of the five components independence covariance model

Component	θ_1	θ_2	θ_3	p_j
1	0.2232	0.0000	25.0084	0.0252
	(0.0976)	(0.0000)	(0.5254)	
2	1.9591	0.1883	8.9511	0.1600
	(0.1523)	(0.0334)	(0.5148)	
3	2.7836	0.3614	0.4916	0.2342
	(0.0471)	(0.0159)	(0.0175)	
4	0.7413	0.4949	2.4109	0.2649
	(0.0412)	(0.0098)	(0.0699)	
5	0.3781	0.0000	0.0085	0.3156
	(0.1096)	(0.0000)	(0.0138)	

Figure 6.5 illustrates the evolution of the loglikelihood for different components (k=1,…,7) of the common covariance multivariate Poisson model. Furthermore, the figure demonstrates that both the AIC and the BIC select five components solution. Figure 6.6 illustrates the optimal value of the mixing proportions for the entire range of models used (values of k from 2 to 7). Again, it can be seen that there is one large component and the rest are small components in all models, except the seven-component model. The mixing proportions tend to fluctuate over the different component solutions.

Table 6.7 contains the parameter estimates for the model with five components. The components with small mixing proportions got the larger estimated standard errors compared to relatively large other components. The parameters with zero estimated values were not differing significantly from zero.

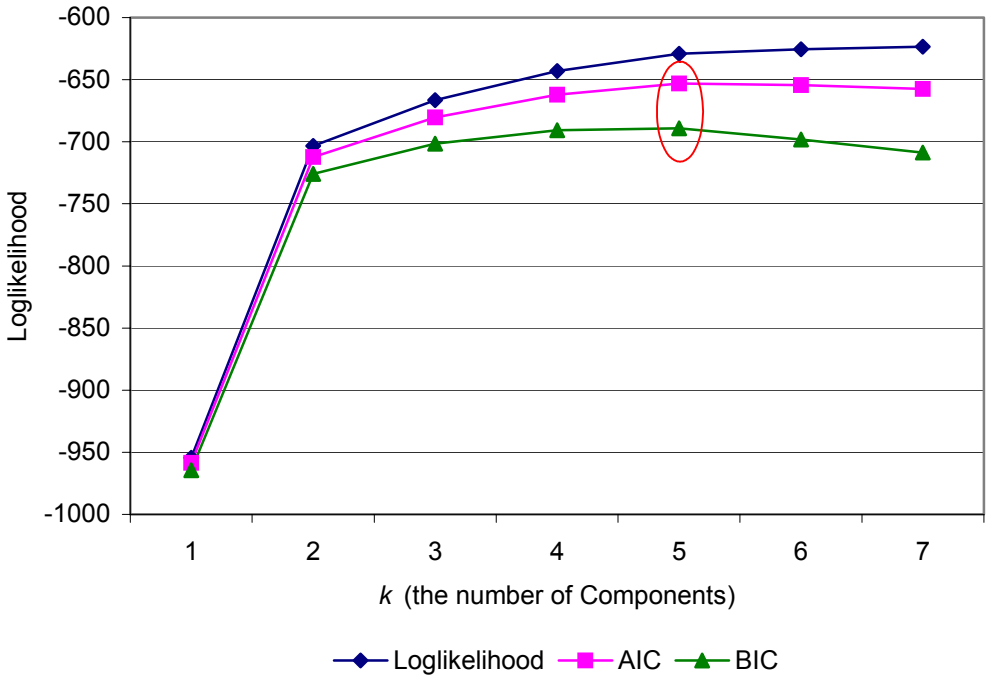


Figure 6.5: Loglikelihood, AIC and BIC against the number of components for the common covariance multivariate Poisson finite mixture model

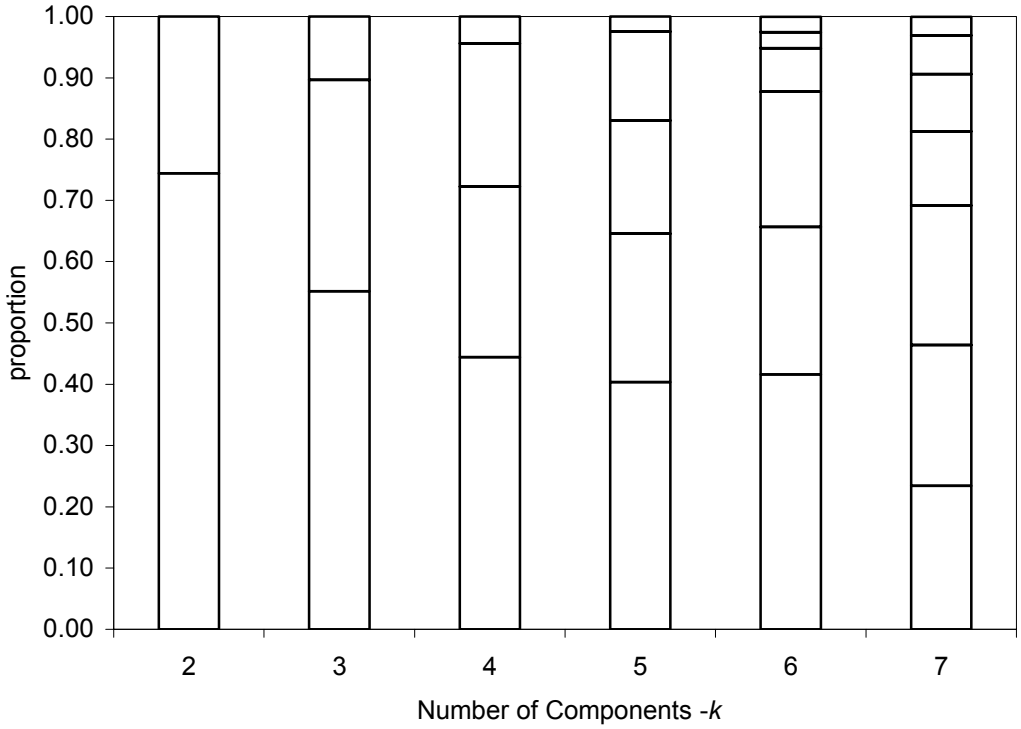


Figure 6.6: The mixing proportions for model solutions with k =2 to 7 components for the common covariance multivariate Poisson finite mixture model

Table 6.7: Parameter estimates (bootstrapped standard errors) of the five components common covariance model

Component	θ_1	θ_2	θ_3	θ_{123}	p_j
1	0.1974	0.0000	25.2688	0.0000	0.0242
	(0.0610)	(0.0385)	(0.6622)	(0.0263)	
2	2.0218	0.1619	9.3893	0.0000	0.1451
	(0.1441)	(0.0191)	(0.1937)	(0.0000)	
3	3.2147	0.2759	0.5463	0.0369	0.1850
	(0.0784)	(0.0143)	(0.0988)	(0.0051)	
4	0.8218	0.4365	2.9303	0.0278	0.2424
	(0.0272)	(0.0056)	(0.0037)	(0.0039)	
5	0.4487	0.1307	0.0494	0.0000	0.4032
	(0.0272)	(0.0056)	(0.0037)	(0.0000)	

Figure 6.7 illustrates the evolution of the loglikelihood for different componets (k =1,…,7) of the restricted covariance multivariate Poisson finite mixture model.

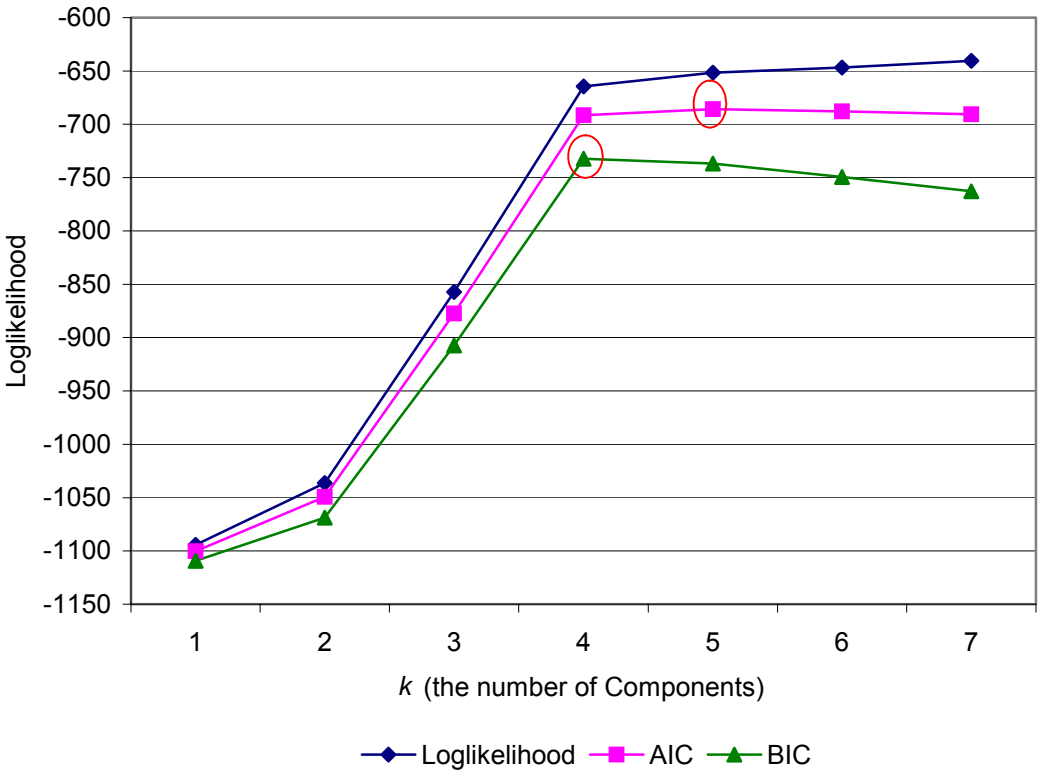


Figure 6.7: Loglikelihood, AIC and BIC against the number of components for the restricted covariance multivariate Poisson finite mixture model

Figure 6.8 illustrates the optimal value of the mixing proportions for the entire range of models used (values of k from 2 to 7).

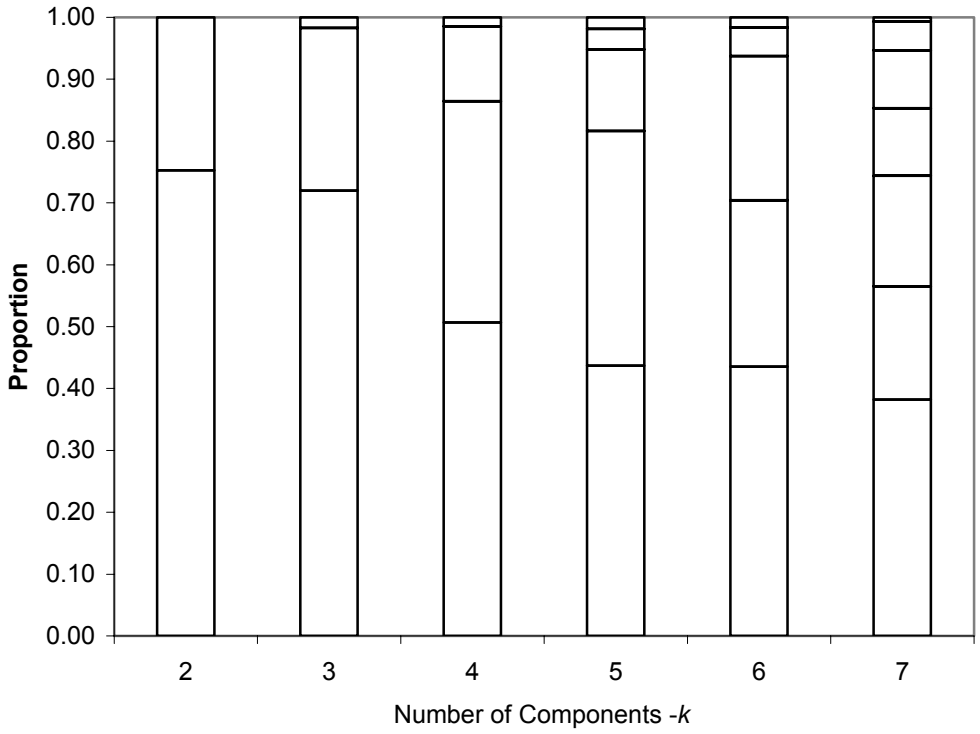


Figure 6.8: The mixing proportions for model solutions with k =2 to 7 components for the restricted covariance multivariate Poisson finite mixture model

Again, the graph does not illustrate a stable cluster configuration, i.e. a clustering that remains relatively stable over the different component solutions. In other words, the cluster proportions tend to fluctuate. It can be seen that there is one large component and the rest are small components in all models. Table 6.8 contains the parameter estimates for the model with five components.

Table 6.8: Parameter estimates (bootstrapped standard errors) of the four component restricted covariance model

Component	θ_1	θ_2	θ_3	θ_{12}	θ_{13}	θ_{23}	p_j
1	6.4384	0.0152	8.5477	0.0000	0.0000	2.4015	0.0143
	(0.3673)	(0.0318)	(0.3180)	(0.0000)	(0.0000)	(0.1516)	
2	0.8485	0.1696	13.5921	0.0000	0.0000	0.0000	0.1213
	(0.0277)	(0.0070)	(0.0778)	(0.0000)	(0.0000)	(0.0000)	
3	1.9083	0.4127	2.8167	0.0000	0.0000	0.0000	0.3575
	(0.0376)	(0.0055)	(0.0285)	(0.0000)	(0.0000)	(0.0000)	
4	0.8075	0.1545	0.0819	0.0000	0.0000	0.0000	0.5069
	(0.0249)	(0.0045)	(0.0049)	(0.0000)	(0.0000)	(0.0000)	

For the restricted covariance model it is also observed that the components of the model with small mixing proportions have the large standard errors.

6.4.2 Results for the different multivariate Poisson hidden Markov models

Similar to the multivariate Poisson finite mixture models, for the multivariate Poisson hidden Markov models all three models, that is, the local independence model, the common covariance model, and the model with restricted covariance structure were fitted sequentially for 1 to 7 components (k =1,…,7). Furthermore, in order to overcome the well-known drawback of the EM algorithm, i.e. the dependence on the initial starting values for the model parameters, 10 different sets of starting values were chosen at random. In fact, the transition probabilities (P_{ij}) were uniform random numbers with constraint $\sum_{j=1}^{m} P_{ij} = 1,\ 1 \le i \le m$. The λ's were generated from a uniform distribution

over the range of the data points. For each set of starting values, the algorithm was run for 200 iterations without caring about any convergence criterion. Then, the set of initial starting values with the largest loglikelihood was selected. The EM iteration were continued with these selected initial values until the convergence criterion is satisfied, i.e. until the relative change of the loglikelihood between two successive iterations was smaller than 10^{-12}. This procedure is repeated 7 times for each value of k.

The selection of number of clusters were based on the most well-known information criterions (section 5.3.4), i.e. the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). The AIC is given as AIC= $L_k - d_k$ and the BIC is given as BIC= $L_k - \ln(n)d_k / 2$ where L_k is the value of the maximized loglikelihood for a model with k components and d_k is the number of free parameters of the model. For the restricted covariance, the independent and the common covariance models d_k is $d_k = 6k + k^2 - k$, $d_k = 3k + k^2 - k$ and $d_k = 4k + k^2 - k$ respectively.

Figure 6.9 illustrates the evolution of the loglikelihood for the different components (k=1,…,7) of the local independence multivariate Poisson hidden Markov model. This figure illustrates that the AIC and the BIC selects five states as the optimal number of states.

Similarly, Figure 6.10 and Figure 6.11 illustrate the evolution of the loglikelihood for the different components of the common covariance and the restricted covariance models respectively. Based on the AIC and the BIC criterion, the five states for the

common covariance and the four states for the restricted covariance model were selected as optimal number of states.

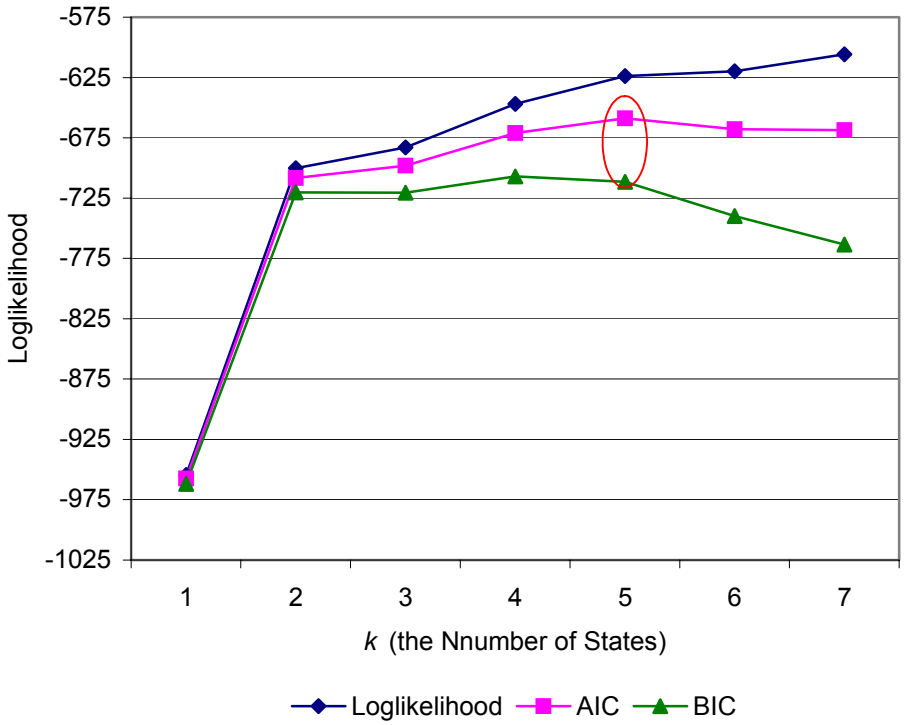


Figure 6.9: Loglikelihood, AIC and BIC against the number of states for the local independent multivariate Poisson hidden Markov model

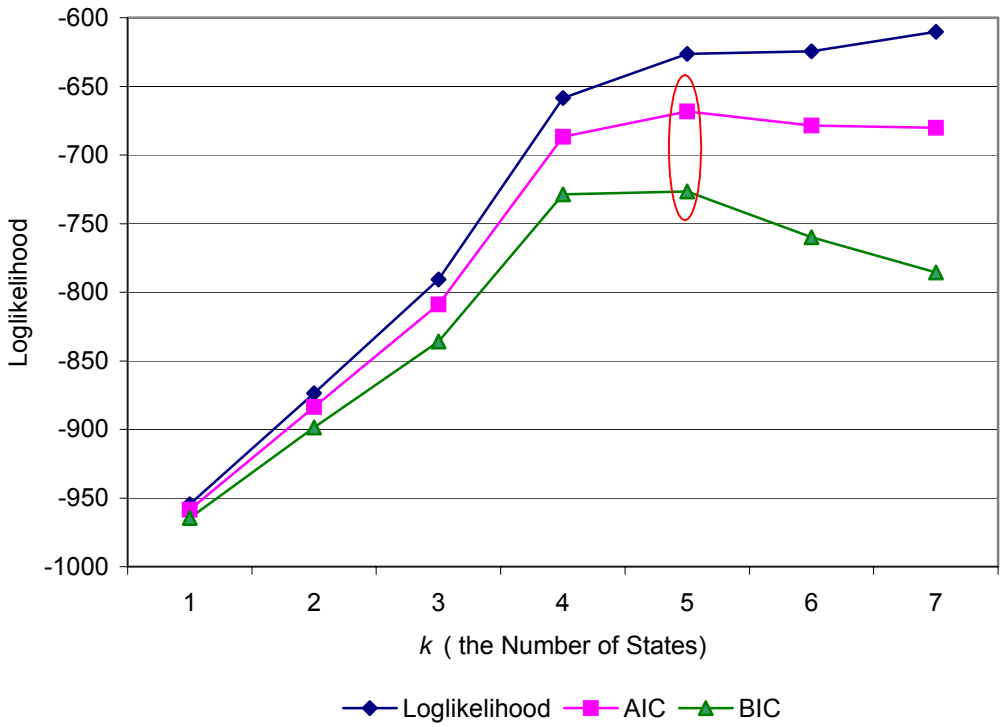


Figure 6.10: Loglikelihood, AIC and BIC against the number of states for the common covariance multivariate Poisson hidden Markov model

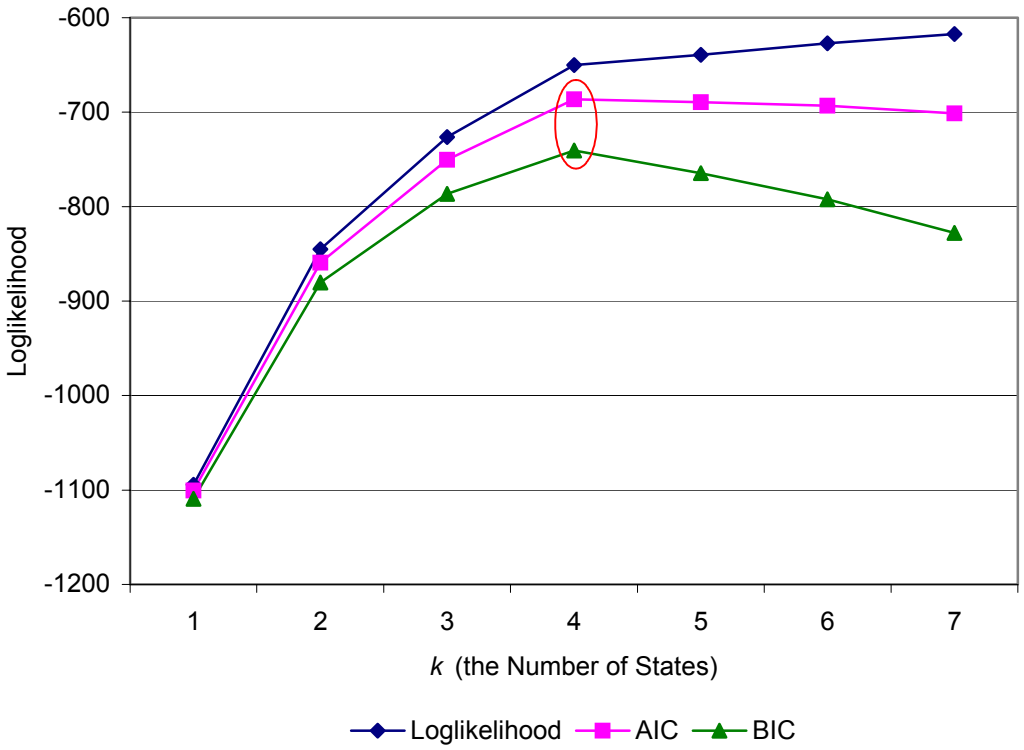


Figure 6.11: Loglikelihood, AIC and BIC against the number of states for the restricted covariance multivariate Poisson hidden Markov model

Table 6.9 contains the parameter estimates and the bootstrapped standard errors for the independent model with five states. The calculation details of the bootstrapped standard errors were given in section 5.5. Here the bootstrap standard errors were considered because of the small sample size (McLachlan et al., 2000), and therefore, the asymptotic standard errors were not valid. Special care was taken to avoid the label switching (Brijs et al., 2004). This problem can be avoided by adding the relevant constraints, $p_1 \leq p_2 \leq ... \leq p_j$ to the optimization algorithm (p_j's are the posterior means of each

state). Parameters with zero estimated values and zero standard errors can be interpreted as zero (Brijs et al., 2004).

Table 6.9: Parameter estimates (bootstrapped standard errors) of the five states hidden Markov independence covariance model

State	λ_1	λ_2	λ_3
1	0.2439	0.0000	24.4691
	(0.1376)	(0.0000)	(0.5506)
2	1.9495	0.2176	8.2134
	(0.0417)	(0.0169)	(0.2339)
3	0.4734	0.5681	2.5843
	(0.0573)	(0.0194)	(0.1480)
4	2.4682	0.3464	0.5381
	(0.0314)	(0.0117)	(0.0318)
5	0.3438	0.0000	0.1149
	(0.0624)	(0.0000)	(0.0415)

Table 6.10: Transition probability matrix of the hidden Markov independence covariance model

$$\begin{bmatrix} 0.4751 & 0.1843 & 0.3406 & 0.0000 & 0.0000 \\ 0.0000 & 0.4260 & 0.0000 & 0.0736 & 0.5003 \\ 0.0000 & 0.0819 & 0.6173 & 0.0455 & 0.2552 \\ 1.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.2187 & 0.1512 & 0.2036 & 0.0000 & 0.4264 \end{bmatrix}$$

Table 6.10 gives the estimated transition probability matrix for the independent model. The $(i, j)^{\text{th}}$ element of the transition matrix is the estimated probability $\widehat{P}_{ij}$ of transition from state state i to state j. It can be seen that some distributions have no chance with probability zero to move to other states. The highest probability of one when moving from state four to state one indicating that the distribution of state four almost surely moved to state one. The next highest probability was 0.5003 when moving from state two to state five.

Table 6.11: Parameter estimates (bootstrapped standard errors) of the five states hidden Markov common covariance model

State	λ_1	λ_2	λ_3	λ_{123}
1	0.2459	0.0000	24.5967	0.0000
	(0.1125)	(0.0194)	(0.8797)	(0.0000)
2	1.9775	0.2086	8.5108	0.0000
	(0.0343)	(0.0089)	(0.1715)	(0.0000)
3	0.5796	0.4304	2.4925	0.0908
	(0.0726)	(0.0124)	(0.1229)	(0.0045)
4	2.0558	0.1685	0.3295	0.0349
	(0.0309)	(0.0066)	(0.0206)	(0.0014)
5	0.0323	0.1043	0.0718	0.0036
	(0.0180)	(0.0116)	(0.0231)	(0.0006)

Table 6.12: Transition probability matrix of the hidden Markov common covariance model

$$\begin{bmatrix} 0.3765 & 0.0778 & 0.1539 & 0.0151 & 0.3767 \\ 0.0000 & 0.0000 & 1.0000 & 0.0000 & 0.0000 \\ 0.1572 & 0.0000 & 0.5334 & 0.0802 & 0.2292 \\ 0.1027 & 0.0668 & 0.0000 & 0.6119 & 0.2186 \\ 0.1372 & 0.0000 & 0.1223 & 0.1583 & 0.5822 \end{bmatrix}$$

Table 6.11 and Table 6.13 contain the parameter estimates and the bootstrapped standard errors for the common covariance and the restricted covariance model with five and four states respectively. Table 6.12 contains the estimated transition probability matrix for the common covariance model. The $(i, j)^{\text{th}}$ element of the transition matrix is the estimated probability $\hat{P}_{ij}$ of transition from state state i to state j. This model had the highest probability of 1 when moving from state two to state three indicating that the distribution of state two almost surely moved to state three. The next highest probability was 0.3767 when moving from state one to state five. Table 6.14 contains

the estimated transition probability matrix for the restricted covariance model. This model had the highest probability of 1 when moving from state four to state three indicating that the distribution of state four almost surely moved to state three. The next highest probability was 0.5383 when moving from state two to state three. It can be seen that the average rates of weed distributions were different in different states. In the restricted covariance model, there is only one important covariance term between Dandelion (Y_2) and Wild Oats (Y_3). Also we see that the distribution of state one only consist of Wild Oats with very high rate (λ_3=41.8286). The state three and state four do not have any correlation between species. The interpretation of other parameters was the same as for the independence case.

Table 6.13: Parameter estimates (bootstrapped standard errors) of the four states hidden Markov restricted covariance model

State	λ_1	λ_2	λ_3	λ_{12}	λ_{13}	λ_{23}
1	0.0000	0.0000	41.8286	0.0000	0.0000	0.0000
	(0.0549)	(0.0000)	(1.5246)	(0.0000)	(0.0000)	(0.0000)
2	1.6504	0.2749	9.6423	0.0000	0.0000	0.7309
	(0.1101)	(0.0473)	(1.6981)	(0.0000)	(0.0000)	(0.1452)
3	1.8772	0.4052	1.8312	0.0000	0.0000	0.0000
	(0.0284)	(0.0204)	(0.3436)	(0.0000)	(0.0000)	(0.0000)
4	0.6173	0.0941	0.0689	0.0000	0.0000	0.0000
	(0.0188)	(0.0115)	(0.0221)	(0.0000)	(0.0000)	(0.0000)

Table 6.14: Transition probability matrix of the hidden Markov restricted covariance model

$$\begin{bmatrix} 0.6733 & 0.1444 & 0.1663 & 0.0160 \\ 0.1438 & 0.3179 & 0.5383 & 0.0000 \\ 0.2883 & 0.1544 & 0.5573 & 0.0000 \\ 0.0000 & 0.0000 & 1.0000 & 0.0000 \end{bmatrix}$$

6.5 Comparison of the different models

Looking at the empirical results of the different model formulations in the previous sections, the following conclusions can be drawn with regard to the fit of the different cluster solutions.

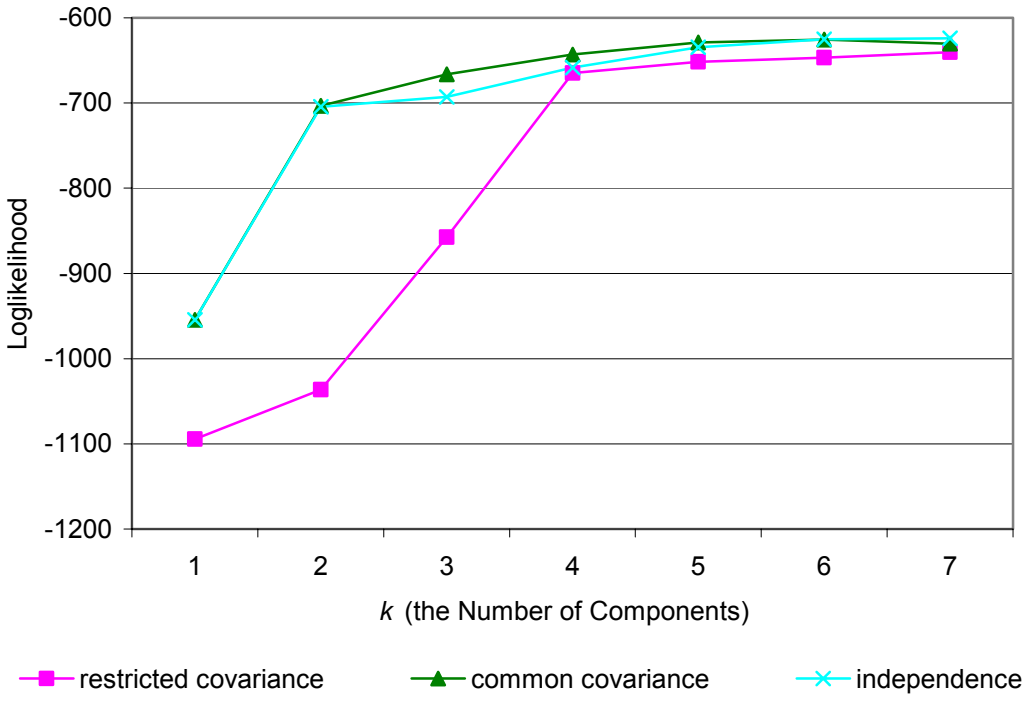


Figure 6.12: Loglikelihood against the number of components (k) for the multivariate Poisson finite mixture models

With regard to the fit of the different models, it was clear from Figure 6.12 for the multivariate Poisson finite mixture models and from Figure 6.13 for the multivariate Poisson hidden Markov models that by adding additional components the fit of the

model, as indicated by the loglikelihood values, increases significantly. Figure 6.12 indeed illustrates that the loglikelihood of the independence and the common covariance models is higher than the loglikelihoods of the restricted covariance model over the range of component solutions (k=1 to 7). Figure 6.13 illustrates that the loglikelihood of the independence model is higher than the loglikelihoods of the restricted and the common covariance model over the range of component solutions (k=1 to 7) for the hidden Markov models.

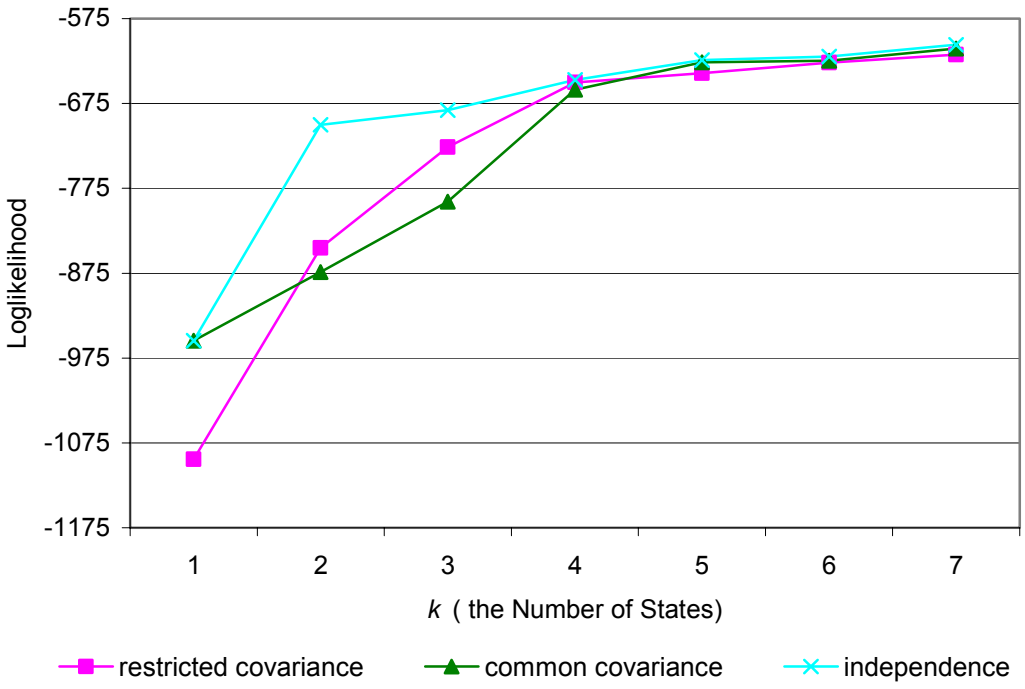


Figure 6.13: Loglikelihood against the number of components (k) for the multivariate Poisson hidden Markov models

From the viewpoint of model fit, Figure 6.13 this partly justifies the use of the model with the independent covariance structure since the comparison of maximized

loglikelihood providing at least a rough indication of the relative goodness of fit. The same conclusion was gained after the primarily loglinear analysis and the correlation matrix of the data.

In order to assess the quality of clustering, the entropy criterion was calculated based on the posterior probabilities (McLachlan et al., 2000 and Brijs et al., 2004). A measure of the strength of clustering is implied by the maximum likelihood estimates in terms of the fitted posterior probabilities of component membership w_{ij} for the finite mixture models and $u_j(i)$ for the hidden Markov models. For example, if the maximum of w_{ij} or $u_j(i)$ is near to 1 for most of the observations, then it suggests that the clusters or states were well separated (McLachlan et al., 2000). The overall measure of strength can be assessed by the average of the maximum of the component-posterior probabilities over the data. The average measure can be represented by the entropy criterion given as

$$I(k) = 1 - \frac{\sum_{i=1}^{n}\sum_{j=1}^{k} w_{ij} \ln(w_{ij})}{n \ln(1/k)}$$

for the finite mixture model with the convention that $w_{ij} \ln(w_{ij}) = 0$ if $w_{ij} = 0$ and

$$I(m) = 1 - \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} u_j(i) \ln(u_j(i))}{n \ln(1/m)}$$

for the hidden Markov model with the convention that $u_j(i)\ln(u_j(i)) = 0$ if $u_j(i) = 0$. In the case of perfect classification, for each *i* there is only one $u_j(i) = 1$ and all the rest are 0 for the hidden Markov model: therefore, the values near to 1 indicate a good

clustering. For our data $I(5)$=0.7686 for the independent model, $I(5)$=0.7837 for the common covariance model and $I(4)$=0.8568 for the restricted covariance model for class of finite mixture models. In a similar manner, for the hidden Markov models, the entropy criterions were $I(5)$=0.8425 for the independent model, $I(5)$=0.8119 for the common covariance model and $I(4)$=0.8441 for the restricted covariance model. Both classes of models indicate that the restricted covariance model had a very good separation between components or states. Among these six models for the finite mixture and the hidden Markov models, the entropy statistic is between 76%-85%. All models can be considered, as "well separated" and the hidden Markov models had a very good separation compared to the finite mixture models.

In the case of the multivariate finite mixture model, each multivariate observation can be allocated to the clusters using the posterior probabilities. The multivariate observation with the highest posterior probability in the k^{th} cluster will be allocated to the k^{th} cluster. Figure 6.14 illustrates the contour plots of clusters for the independent, the common and the restricted covariance multivariate finite mixture models.

Given the sequence of observations $\mathbf{Y}$ and the model with the transition probability matrix, the most likely state sequence associated with the given observation sequence can be found. This can be achieved by maximizing the probability of observing observation sequence and the state sequence given their joint distribution. This can be achieved using the so-called Viterbi Algorithm (Viterbi, 1967). After allocating each observation to the corresponding states the optimal state path can be found. Figure 6.15

illustrates the contour plot of the independent, the common and the restricted covariance hidden Markov models, which visualized the pattern of the weed distributions.

Comparing Figure 6.14 and Figure 6.15 it can be seen that there were similarities in the weed distributions from the finite mixture model allocation and the hidden Markov model allocation for the three covariance structures. For the restricted covariance model, the allocation of observations to the clusters or states was very similar for both models. But for the independent and the common covariance structures the allocation of some of the observations to clusters or states was not the same.

The choice of a best model is still questionable. In the next chapter, properties of the finite mixture models and a criterion for goodness of fit index are discussed.

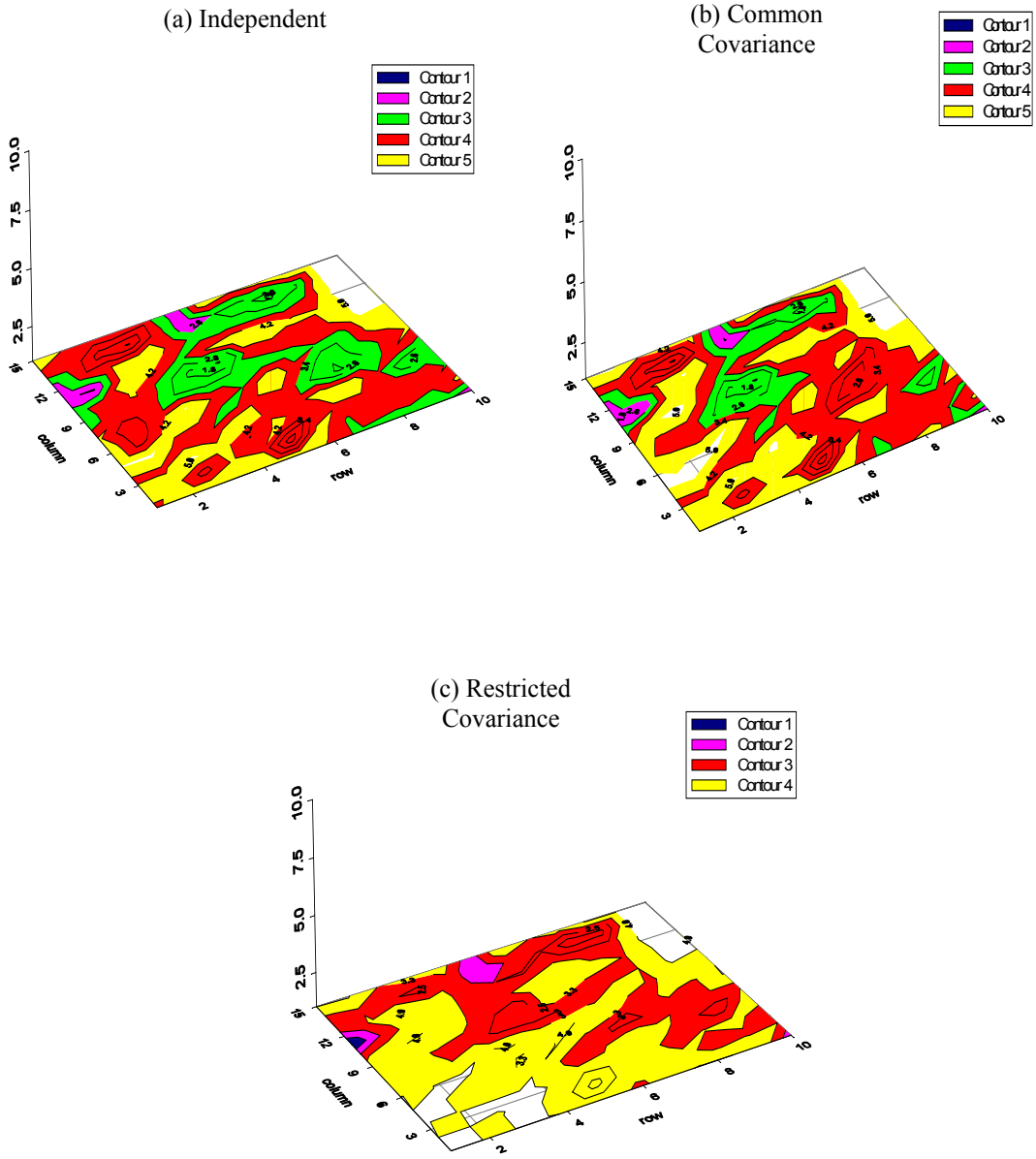


Figure 6.14: Contour plots of clusters for the (a) independent, (b) common and (c) restricted covariance multivariate finite mixture models.

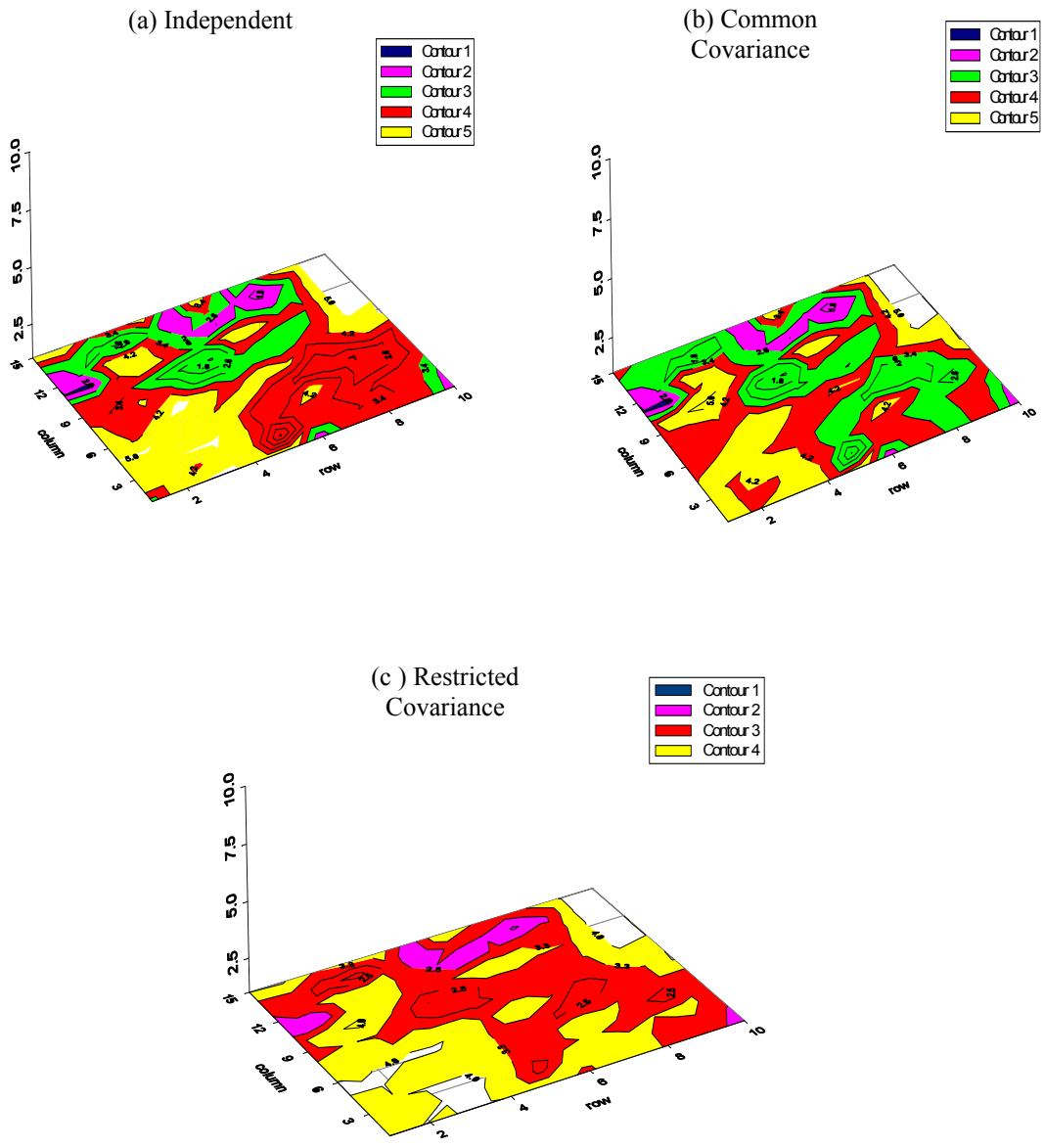


Figure 6.15: Contour plots of clusters for the (a) independent, (b) common and (c) restricted covariance multivariate Poisson hidden Markov models

CHAPTER 7

PROPERTIES OF MULTIVARIATE POISSON FINITE MIXTURE MODELS AND APPLICATIONS

7.1 Introduction

In this chapter, the properties of the multivariate Poisson finite mixture models are discussed. The importance of exploring the properties of the finite mixtures, the extension of these properties to the hidden Markov model and the application to other data sets are presented in the next sections.

Even though there was more literature available on the analysis of count data, still only small portions of it deal with correlated counts. Holgate (1964) discussed the estimation problems of the bivariate Poisson distribution which does not support negative correlation between the two count variables. With the availability of powerful computing facilities Aitchison and Ho (1989) described how the multivariate lognormal mixture of the independent Poisson distributions could take into account the positive and negative correlation between the variables. A class of models proposed by Chib and Winkelmann (2001) can take into account the correlation among the counts. They developed an efficient Markov Chain Monte Carlo algorithm to estimate the model parameters. However, for these models, the computational burden was quite large.

Karlis and Meligkotsidou (2006) discussed the correlation structure of the multivariate Poisson mixture models. These mixture models allow for both negative correlations and overdispersion in addition to being computationally feasible.

The multivariate Poisson distribution is discussed again in section 7.2, followed by the properties of the finite mixture models. In section 7.5, these properties were applied to both multivariate Poisson finite mixture models and multivariate Poisson hidden Markov models for several applications.

7.2 The multivariate Poisson distribution

Consider a vector $\mathbf{X} = (X_1, X_2, ..., X_k)$ where X_i's are independent and each follows a Poisson distribution with parameter $\lambda_j, j = 1,...,k$. Suppose that matrix $\mathbf{A}$ has dimensions $n \times k$ with zeros and ones. Then the vector $\mathbf{Y} = (Y_1, Y_2, ..., Y_n)$ defined as the $\mathbf{Y} = \mathbf{AX}$ follows a *n*-variate Poisson distribution. The most general form of a *n*-variate Poisson distribution assumes that **A** is a matrix of size $n \times (2^n - 1)$ of the form

$$\mathbf{A}=[\mathrm{A}_1, \mathrm{A}_2, \mathrm{A}_3,\ldots,\mathrm{A}_n]$$

where $\mathbf{A}_i$, $i = 1,...,n$ are matrices with n rows and $\binom{n}{i}$ columns. The matrix $\mathbf{A}_i$ contains columns with exactly i ones and $(n-i)$ zeros, with no duplicate columns, for $i = 1,...,n$. Thus $\mathbf{A}_n$ is the column vector of 1's, while $\mathbf{A}_1$ becomes the identity matrix of size $n \times n$. For example, the fully structured multivariate Poisson model for three variables can be represented as follows:

$$Y_1 = X_1 + X_{12} + X_{13} + X_{123}$$
$$Y_2 = X_2 + X_{12} + X_{23} + X_{123}$$
$$Y_3 = X_3 + X_{13} + X_{23} + X_{123}$$

$$\mathbf{Y} = \mathbf{AX}$$

$$\mathbf{A}=[\mathrm{A}_1, \mathrm{A}_2, \mathrm{A}_3]$$

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & 1 \end{bmatrix}$$

$$\text{where } \mathbf{A}_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \mathbf{A}_2 = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix}, \mathbf{A}_3 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

$$\text{and } \mathbf{X} = (X_1, X_2, X_3, X_{12}, X_{13}, X_{23}, X_{123}).$$

The reduced models for *n* variables can be derived from selecting the $\mathbf{A}$ matrix. The restricted covariance trivariate Poisson model can be presented as follows:

$$Y_1 = X_1 + X_{12} + X_{13}$$
$$Y_2 = X_2 + X_{12} + X_{23}$$
$$Y_3 = X_3 + X_{13} + X_{23}$$

$$\mathbf{A}=[\mathrm{A}_1, \mathrm{A}_2]$$

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 \end{bmatrix}$$

$$\text{where } \mathbf{A}_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \mathbf{A}_2 = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix}.$$

The mean vector and the covariance matrix of the vector $\mathbf{Y}$ are given as:

$$E(\mathbf{Y}) = \mathbf{AM} \text{ and } Var(\mathbf{Y}) = \mathbf{A\Sigma}A^{\mathrm{T}}, \tag{7.1}$$

where $\boldsymbol{M}$ is the mean vector of $\mathbf{X}$ and is given as: $\mathbf{M} = E(\mathbf{X}) = (\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, ..., \boldsymbol{\lambda}_k)^T$ and $\boldsymbol{\Sigma}$ is the variance and covariance matrix of $\mathbf{X}$ and is given as: $\boldsymbol{\Sigma} = Var(\mathbf{X}) = diag(\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, ..., \boldsymbol{\lambda}_k)$. Since $\mathbf{X}$'s are independent, $\boldsymbol{\Sigma}$ is diagonal matrix. More details and references for the multivariate Poisson model can be found in Karlis and Xekalaki (2005). The identifiability and the consistency of finite mixtures of the multivariate Poisson distribution with two-way covariance structure are proved in Karlis and Meligkotsidou (2006).

In general notation, let $f(y;\boldsymbol{\lambda}) \bigwedge_{\boldsymbol{\lambda}} g(\boldsymbol{\lambda})$ be a general mixture of the density $f(y;.)$ with respect to its parameter $\boldsymbol{\lambda}$, where $g(\boldsymbol{\lambda}), \boldsymbol{\lambda} \in \boldsymbol{\Theta}$ is the mixing distribution. The density of the mixing distribution is given by $f(y) = \int_{\boldsymbol{\Theta}} f(y;\boldsymbol{\lambda}) dG(\boldsymbol{\lambda})$, where $G(\boldsymbol{\lambda})$ is the cumulative function of the mixing distribution.

7.3 The properties of finite mixture models

The joint probability function of $\boldsymbol{Y}$ is $p(\mathbf{y};\boldsymbol{\lambda})$, and then the finite multivariate Poisson mixture distribution can be given as:

$$f(\mathbf{y}) = \sum_{j=1}^{k} p_j p(\mathbf{y};\boldsymbol{\lambda}_j) \tag{7.2}$$

where p_j's are mixing proportions and the marginal distributions are finite mixtures. Then the expectation of the finite multivariate Poisson mixture is given as:

$$E(\mathbf{Y}) = \sum_{j=1}^{k} p_j \mathbf{A}\mathbf{M}_j \ , \tag{7.3}$$

$$\text{where } \mathbf{M}_j = \boldsymbol{\lambda}_{tj}^T; t \in \boldsymbol{\Omega}, \boldsymbol{\Omega} = \{1,2,3,12,13,23\}$$

for the reduced model.

Different covariance structures can be formed for the different subpopulations by changing the matrix ***A*** for each subpopulation.

Recalling that the covariance of ***X*** conditional on the vector $\boldsymbol{\lambda}$ (Karlis and Meligkotsidou, 2006)

$$\boldsymbol{\Sigma} = Var(\mathbf{X} \mid \boldsymbol{\lambda}) = \begin{bmatrix} \lambda_1 & 0 & ... & 0 \\ 0 & \lambda_2 & ... & 0 \\ 0 & .... & .... & 0 \\ 0 & .... & .... & \lambda_t \end{bmatrix}, \tag{7.4}$$

where ***X*** denotes the vector of the latent variables used to construct the multivariate Poisson distribution. The second moment of ***Y*** conditional on $\boldsymbol{\lambda}$ is given by

$$\begin{aligned} E(\mathbf{Y}\mathbf{Y}^T \mid \boldsymbol{\lambda}) &= Var(\mathbf{Y} \mid \boldsymbol{\lambda}) + E(\mathbf{Y} \mid \boldsymbol{\lambda})[E(\mathbf{Y} \mid \boldsymbol{\lambda})]^T \\ &= \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T + \mathbf{A}\mathbf{M}(\mathbf{A}\mathbf{M})^T \\ &= \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T + \mathbf{A}\mathbf{M}\mathbf{M}^T\mathbf{A}^T \\ &= \mathbf{A}[\boldsymbol{\Sigma} + \mathbf{M}\mathbf{M}^T]\mathbf{A}^T \end{aligned} \tag{7.5}$$

Let $\mathbf{B}(\boldsymbol{\lambda}) = \boldsymbol{\Sigma} + \mathbf{M}\mathbf{M}^T$ and $\mathbf{B}(\boldsymbol{\lambda})$ has the following form:

$$\mathbf{B}(\boldsymbol{\lambda}) = \begin{bmatrix} \lambda_1^2 + \lambda_1 & \lambda_1\lambda_2 & ... & \lambda_1\lambda_t \\ \lambda_1\lambda_2 & \lambda_2^2 + \lambda_2 & ... & \lambda_2\lambda_t \\ ... & ... & ... & ... \\ \lambda_1\lambda_t & ... & ... & \lambda_t^2 + \lambda_t \end{bmatrix}. \tag{7.6}$$

The simple moments of **B** are polynomial with respect to the parameters $\lambda_t, t \in \{1,2,3,12,13,23\}$, and thus, the moments of the mixture can be obtained as functions of the moments of the mixing distribution G using the standard expectation argument given below.

$$E(Y_i^r Y_j^s) = \int E(Y_i^r Y_j^s \mid \lambda) dG(\lambda), \tag{7.7}$$

where *r,s* = 0,1,…. The element-wise expectations of a matrix $\mathbf{B}(\boldsymbol{\lambda})$ can be represented as:

$$E(\mathbf{B}) = \begin{bmatrix} E(\lambda_1^2) + E(\lambda_1) & E(\lambda_1\lambda_2) & \ldots & E(\lambda_1\lambda_t) \\ E(\lambda_1\lambda_2) & E(\lambda_2^2) + E(\lambda_2) & \ldots & E(\lambda_2\lambda_t) \\ & \ldots & & \\ E(\lambda_1\lambda_t) & \ldots & & E(\lambda_t^2) + E(\lambda_t) \end{bmatrix}, \tag{7.8}$$

then the unconditional variance of the vector **Y** (Karlis and Meligkotsidou, 2006), that is the variance covariance matrix of the mixture is

$$Var(\mathbf{Y}) = E(\mathbf{Y}\mathbf{Y}^T) - E(\mathbf{Y})[E(\mathbf{Y})]^T, \text{ where } E(\mathbf{Y}\mathbf{Y}^T) = \mathbf{A}E(\mathbf{B})\mathbf{A}^T. \tag{7.9}$$

The moments of the multivariate Poisson distribution are simple polynomials with respect to the mixing parameters. Comparing the estimated unconditional covariance matrix to its observed covariance matrix can be used as a goodness of fit index.

For example, consider the three components trivariate finite mixture model with following restricted covariance structure.

$$Y_1 = X_1 + X_{12} + X_{13}$$
$$Y_2 = X_2 + X_{12} + X_{23}$$
$$Y_3 = X_3 + X_{13} + X_{23}$$

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 \end{bmatrix} \text{ and}$$

p_1, p_2 and p_3 are mixing proportions. The element wise expectation of matrix $\mathbf{B}(\boldsymbol{\lambda})$ are $E(\mathbf{B}_1)$, $E(\mathbf{B}_2)$ and $E(\mathbf{B}_3)$ respectively. Then

$$E(\mathbf{B}) = p_1 E(\mathbf{B}_1) + p_2 E(\mathbf{B}_2) + p_3 E(\mathbf{B}_3)$$

$$E(\mathbf{B}_1) = \begin{bmatrix} E(\lambda_{11}^2) + E(\lambda_{11}) & E(\lambda_{11}\lambda_{21}) & \ldots & E(\lambda_{11}\lambda_{t1}) \\ E(\lambda_{11}\lambda_{21}) & E(\lambda_{21}^2) + E(\lambda_{21}) & \ldots & E(\lambda_{21}\lambda_{t1}) \\ & \ldots & & \\ E(\lambda_{11}\lambda_{t1}) & \ldots & & E(\lambda_{t1}^2) + E(\lambda_{t1}) \end{bmatrix}$$

$$E(\mathbf{B}_2) = \begin{bmatrix} E(\lambda_{12}^2) + E(\lambda_{12}) & E(\lambda_{12}\lambda_{22}) & \ldots & E(\lambda_{12}\lambda_{t2}) \\ E(\lambda_{12}\lambda_{22}) & E(\lambda_{22}^2) + E(\lambda_{22}) & \ldots & E(\lambda_{22}\lambda_{t2}) \\ & \ldots & & \\ E(\lambda_{12}\lambda_{t2}) & \ldots & & E(\lambda_{t2}^2) + E(\lambda_{t2}) \end{bmatrix}$$

$$E(\mathbf{B}_3) = \begin{bmatrix} E(\lambda_{13}^2) + E(\lambda_{13}) & E(\lambda_{13}\lambda_{23}) & \ldots & E(\lambda_{13}\lambda_{t3}) \\ E(\lambda_{13}\lambda_{23}) & E(\lambda_{23}^2) + E(\lambda_{23}) & \ldots & E(\lambda_{23}\lambda_{t3}) \\ & \ldots & & \\ E(\lambda_{13}\lambda_{t3}) & \ldots & & E(\lambda_{t3}^2) + E(\lambda_{t3}) \end{bmatrix}$$

and $E(\mathbf{Y}) = \mathbf{AM}$ where $\mathbf{M} = \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \cdots \\ \lambda_t \end{bmatrix}$.

Details of the proof of the unconditional variance of vector $\mathbf{Y}$ are given in Karlis and Meligkotsidou (2006). A brief description of the multivariate Poisson-log normal distribution is given in the next section, and these models were compared with the finite mixture models in section 7.5.

7.4 Multivariate Poisson-log Normal distribution

The multivariate Poisson-log normal distribution (Aitchison and Ho, 1989) is a natural extension of the univariate Poisson-log normal distribution. Here the mixing of *d* independent Poisson distributions $Po(\lambda_i)$ is achieved by placing a *d*-dimensional lognormal distribution on the *d*-dimensional vector $\boldsymbol{\lambda}$. The multivariate Poisson-log normal distribution supports negative and positive correlation between the count variables.

7.4.1 Definition and the properties

Let $g(\boldsymbol{\lambda} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denote the probability density function of the *d*-dimensional log normal distribution $\boldsymbol{\Lambda}^d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, so that

$$g(\boldsymbol{\lambda} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-\frac{1}{2}d} (\lambda_1, ..., \lambda_d)^{-1} \mid \Sigma \mid^{-\frac{1}{2}} \exp\{-\frac{1}{2}(\log \lambda - \mu)^{-1} \Sigma^{-1} (\log \lambda - \mu)\} \qquad (7.10)$$

The multivariate Poisson-log normal distribution denoted by $\mathbf{P\Lambda}^d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the $\boldsymbol{\Lambda}^d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ mixture of independent $Po(\lambda_i)$ distributions (i =1,…, d) with probability density function

$$p(\mathbf{y} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \int_{R_+^d} \prod_{i=1}^{d} f(\mathbf{y}_i \mid \boldsymbol{\lambda}_i) g(\boldsymbol{\lambda} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) d\boldsymbol{\lambda}\,;\ (y_1, ..., y_d = 0,1,...) \qquad (7.11)$$

where R_+^d denotes the positive orthant of *d*-dimensional real space R^d.

It is not easy to simplify the multiple integral (7.11), but its moments can be easily obtained through conditional expectation results and standard properties of the Poisson and log normal distributions. The expectation, variance, covariance, and correlation for the multivariate Poisson-log normal model are given below (Aitchison and Ho, 1989).

Let σ_{ij} denotes the (*i, j*) element of $\boldsymbol{\Sigma}$. Then

$$E(Y_i) = \exp(\mu_i + \frac{1}{2}\sigma_{ii}) = \alpha_i. \qquad (7.12)$$

$$Var(Y_i) = \alpha_i + \alpha_i^2 \{\exp(\sigma_{ii}) - 1\}. \qquad (7.13)$$

$$Cov(Y_i, Y_j) = \alpha_i \alpha_j \{\exp(\sigma_{ii}) - 1\}. \qquad (7.14)$$

$$Corr(Y_i, Y_j) = \frac{\exp(\sigma_{ij}) - 1}{[\{\exp(\sigma_{ii}) - 1 + \alpha_i^{-1}\}\{\exp(\sigma_{jj}) - 1 + \alpha_j^{-1}\}]^{\frac{1}{2}}}. \qquad (7.15)$$

7.5 Applications

In addition to weed species data, two other data sets (The lens faults dataset p.649 and the bacterial count data set p.651) presented in Aitchison and Ho (1989) were used to compare the models among the multivariate Poisson-log normal distribution, the multivariate Poisson finite mixture and the hidden Markov model (Markov-dependent finite mixture model). Calculations were carried out for the hidden Markov model, replacing the mixing proportions in finite mixtures by posterior means of each state. These posterior means were used to assess the goodness of fit of the hidden Markov model (HMM). Results are presented and discussed in the next section.

7.5.1 The lens faults data

Table 7.1: Counts (x_1, x_2) of surface and interior faults in 100 lenses

				x_2								
x_1	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**9**	**10**	**12**	**14**
0	1	1	4									1
1	3	2	6	2	5		2					
2	1	2	4	3	2	1	1	1	1		1	
3		5	1	2	2	3	2					
4	1	2	2	5	3	1	1					
5	1	2	1	2	1	2				1		
6	2	2		1	1			1				
7	1	3		1								
8		2										
11		1										
12		1										

Table 7.1 gives the counts of surface and interior faults in 100 lenses presented in Aitchison and Ho (1989). The observed covariance matrix and the correlation between surface (x_1) and interior (x_2) counts are given below:

$$\begin{bmatrix} 5.2227 & -1.0227 \\ -1.0227 & 6.2072 \end{bmatrix} \quad r = -0.1796.$$

The correlation coefficient indicates that data have a negative correlation. Table 7.2 gives the loglikelihood, the AIC and the BIC together with the number of components for the bivariate common covariance Poisson finite mixture model (7.16).

$$p(x_1, x_2) = \sum_{j=1}^{k} p_j Po(x_1, x_2 | \lambda_{1j}, \lambda_{2j}, \lambda_{3j}),$$

$$\text{where } Po(x_1, x_2 \mid \lambda_1, \lambda_2, \lambda_3) = e^{-(\lambda_1+\lambda_2+\lambda_3)} \frac{\lambda_1^{x_1}}{x_1!} \frac{\lambda_2^{x_2}}{x_2!} \sum_{i=0}^{\min(x_1, x_2)} \binom{x_1}{i} \binom{x_2}{i} i! \left(\frac{\lambda_3}{\lambda_1 \lambda_2} \right)^i. \qquad (7.16)$$

According to the AIC and the BIC criterion (section 5.3.4), the larger the criterion, the better the model in comparison with another. Therefore, the three-component model with loglikelihood –420.6121 was selected as the best model (Table 7.2). The covariance matrix and the correlation between x_1 and x_2 were estimated.

Table 7.2: Loglikelihood, AIC and BIC together with the number of components for the common covariance multivariate Poisson finite mixture Model

Number of components (k)	Number of free parameters	Loglikelihood	AIC	BIC
1	3	-450.6038	-453.6038	-457.5115
2	7	-432.6901	-439.6901	-448.8082
3	11	-420.6121	-431.6121	-445.9405
4	15	-419.8284	-434.8284	-454.3672
5	19	-419.7168	-438.7168	-463.4659
6	23	-419.3221	-442.3221	-472.2815
7	27	-419.3221	-446.3221	-481.4919

The estimated covariance matrix and correlation coefficient are given below:

$$\begin{bmatrix} 5.3742 & -0.9564 \\ -0.9564 & 6.1084 \end{bmatrix} \quad r = -0.1669.$$

Table 7.3 gives the logliklihood, the AIC and the BIC together with the number of components for the bivariate independent Poisson finite mixture model (7.17).

$$p(x_1, x_2) = \sum_{j=1}^{k} p_j Po(x_1, x_2 | \lambda_{1j}, \lambda_{2j}),$$

$$\text{where } Po(x_1, x_2 \mid \lambda_1, \lambda_2) = e^{-(\lambda_1+\lambda_2)} \frac{\lambda_1^{x_1}}{x_1!} \frac{\lambda_2^{x_2}}{x_2!} \sum_{i=0}^{\min(x_1, x_2)} \binom{x_1}{i} \binom{x_2}{i} i! . \tag{7.17}$$

In this case, the AIC and the BIC criterion select different component models: the AIC selects the four-component model and the BIC selects the three-component model. The method described in section 7.2 is used to calculate the covariance matrices.

Table 7.3: Loglikelihood, AIC and BIC together with the number of components for the local independence multivariate Poisson finite mixture Model

Number of components (k)	Number of free parameters	Loglikelihood	AIC	BIC
1	2	-450.6038	-452.6038	-455.2089
2	5	-433.5880	-438.5881	-445.1009
3	8	-423.6535	-431.6536	-442.0742
4	11	-420.2611	-431.2615	-445.5895
5	14	-419.2967	-433.2967	-451.5329
6	17	-419.2967	-436.2967	-458.4406

The estimated covariance matrix (AIC selection) and the correlation coefficient are

$\begin{bmatrix} 5.7531 & -1.0169 \\ -1.0169 & 6.3774 \end{bmatrix}$ and $r = -0.1679$ respectively.

The estimated covariance matrix (BIC selection) and the correlation coefficient are

$\begin{bmatrix} 5.2607 & -1.3547 \\ -1.3547 & 6.0806 \end{bmatrix}$ and $r = -0.2395$ respectively.

Table 7.4: Loglikelihood, AIC and BIC together with the number of components for the common covariance multivariate Poisson hidden Markov Model

Number of components (k)	Number of free parameters	Loglikelihood	AIC	BIC
1	3	-450.6038	-453.6038	-457.5115
2	8	-398.6838	-406.6838	-417.1045
3	15	-377.4649	-392.4649	-412.0036
4	24	-361.7565	-385.7565	-417.0185
5	35	-348.8014	-383.8014	-429.3918
6	48	-340.3936	-388.3936	-450.9177

Similarly, the loglikelihood, the AIC and the BIC values for the common covariance and the independent model for the Markov-dependent bivariate Poisson finite mixture models are given in Table 7.4 and Table 7.5, respectively. The corresponding estimated covariance matrices and the correlation coefficients between x_1 and x_2 are also presented.

The estimated covariance matrix (AIC selection) and the correlation coefficient are

$\begin{bmatrix} 6.7930 & -0.3909 \\ -0.3909 & 6.4713 \end{bmatrix}$ and r = -0.0590 respectively.

The estimated covariance matrix (BIC selection) and the correlation coefficient are

$\begin{bmatrix} 6.6645 & -0.6649 \\ -0.6649 & 5.0389 \end{bmatrix}$ and r = -0.1147 respectively.

Table 7.5: Loglikelihood, AIC and BIC together with the number of components for the local independence multivariate Poisson hidden Markov Model

Number of components (k)	Number of free parameters	Loglikelihood	AIC	BIC
1	2	-450.6038	-452.6038	-455.2089
2	6	-398.7847	-404.7847	-412.6002
3	12	-377.4630	-389.4630	-405.0940
4	20	-368.0097	-388.0097	-414.0614
5	30	-359.2484	-389.2484	-428.3259
6	42	-350.8904	-392.8904	-447.5989

The estimated covariance matrix (AIC selection) and the correlation coefficient are

$\begin{bmatrix} 7.3839 & -1.1433 \\ -1.1433 & 5.6561 \end{bmatrix}$ and r = -0.1769 respectively.

The estimated covariance matrix (BIC selection) and the correlation coefficient are

$\begin{bmatrix} 6.6645 & -0.6649 \\ -0.6649 & 5.0389 \end{bmatrix}$ and r = -0.1147 respectively.

Comparing all these models for the bivariate Poisson finite mixture and the hidden Markov models, this goodness of fit criterion suggest that the local independence bivariate Poisson finite mixture model is the best fitted model with respect to the estimated covariance and the correlation parameters.

The observed correlation for lenses count data given in Table 7.1 was -0.1796. Using the multivariate Poisson-log normal distribution Aitchison and Ho (1989) found that the best model with loglikelihood –426.40 and the estimated count correlation between x_1 and x_2 was –0.21. Multivariate Poisson finite mixtures provide a better fit compared to Aitchison and Ho models, having larger loglikelihoods (-420.6121, -420.2611, and -423.6535). It also demonstrates that estimated correlation coefficient is very much close to the observed correlations (AIC selections) except the hidden Markov common covariance model, compared to Aitchison and Ho models. Note that the maximum likelihood estimation of the parameters of the multivariate Poisson-log normal model was obtained by the combination of the Newton-Raphson and the steepest ascent method (Aitchison and Ho, 1989).

7.5.2 The bacterial count data

In the study of relative effectiveness of three different air samplers 1, 2, and 3 to detect pathogenic bacteria in 'sterile' rooms, a microbiologist obtained triplets of bacterial colony counts X_1, X_2, and X_3 from samplers 1, 2, and 3 in each of 50 different sterile locations. Since the bacterial infestation can vary from location to location, extra-Poisson variations can be expected in the counts from any particular sampler, with correlation between the three counts from a particular location. Aitchison and Ho (1989) considered $P\Lambda^3(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ model seems a reasonable framework for these data. They observed maximum loglikelihood –397.8 for $P\Lambda^3(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ model.

Table 7.6 Bacterial counts by 3 samplers in 50 sterile locations

X_1	X_2	X_3	X_1	X_2	X_3	X_1	X_2	X_3	X_1	X_2	X_3	X_1	X_2	X_3
1	2	11	3	6	6	3	8	2	7	10	5	22	9	6
8	6	0	3	9	14	1	1	30	2	2	8	5	2	4
2	13	5	4	2	25	4	5	15	3	15	3	2	0	6
2	8	1	9	7	3	7	6	3	1	8	2	2	1	1
5	6	5	5	4	8	8	10	4	4	6	0	4	6	4
14	1	7	4	4	7	3	2	10	8	7	3	4	9	2
3	9	2	7	3	2	6	8	5	6	6	6	8	4	6
7	6	8	1	14	6	2	3	10	4	14	7	3	10	6
3	4	12	2	13	0	1	7	3	3	3	14	4	7	10
1	9	7	14	9	5	2	9	12	6	8	3	2	4	6

The observed covariance matrix and the observed correlation matrix are given below:

$$\begin{bmatrix} 15.0714 & 0.2755 & -3.6939 \\ 0.2755 & 13.6428 & -7.7347 \\ -3.6939 & -7.7347 & 32.6122 \end{bmatrix} \qquad \begin{bmatrix} 1 & 0.0192 & -0.1666 \\ & 1 & -0.3667 \\ & & 1 \end{bmatrix}.$$

Since the correlations were not very high, only the independence covariance structure is considered and the loglikelihood and the AIC values are recorded. According to the AIC criterion (Table 7.7), the model with the highest AIC value is selected (loglikelihood of –382.2335). This model, the seven-component local independence multivariate Poisson finite mixture model, gives a reasonable fit with respect to the loglikelihood.

Table 7.7: Loglikelihood and AIC together with the number of components for the local independence multivariate Poisson finite mixture Model

Number of components (k)	Number of free parameters	Loglikelihood	AIC
1	3	-472.8759	-475.8759
2	7	-422.3424	-429.3424
3	11	-409.9689	-420.9689
4	15	-402.9517	-417.9517
5	19	-395.5478	-414.5478
6	23	-390.3447	-413.3447
7	27	-382.2335	-409.2335
8	31	-381.3817	-412.3817
9	35	-381.3717	-416.3817

The estimated covariance matrix and the estimated correlation matrix are given below:

$$\begin{bmatrix} 15.9583 & -0.2698 & -3.3550 \\ -0.2698 & 13.3753 & -7.9948 \\ -3.3550 & -7.9948 & 33.4182 \end{bmatrix} \qquad \begin{bmatrix} 1 & -0.0185 & -0.1453 \\ & 1 & -0.3781 \\ & & 1 \end{bmatrix}.$$

The covariance between X_1 and X_2 samplers does not seem to be in the right direction; however, other parameters are close to the observed covariance matrix.

Similar analyses were carried out with our proposed model (local independence multivariate Poisson hidden Markov model) and the results are given in Table 7.8. According to the AIC criterion, the model with five components (loglikelihood – 382.9375) gives a better fit compared to other component models.

Table 7.8: Loglikelihood and AIC together with the number of components for the local independence multivariate Poisson hidden Markov Model

Number of components (k)	Number of free parameters	Loglikelihood	AIC
1	3	-472.8759	-475.8759
2	8	-421.5801	-429.5801
3	15	-406.6949	-421.6949
4	24	-395.4245	-419.4245
5	35	-382.9375	-417.9375
6	48	-375.2770	-423.2770
7	63	-372.5788	-435.5788

The estimated covariance matrix and the estimated correlation matrix are given below:

$$\begin{bmatrix} 14.0247 & 1.4131 & -3.1730 \\ 1.4131 & 11.5828 & -8.2055 \\ -3.1730 & -8.2055 & 32.9665 \end{bmatrix} \qquad \begin{bmatrix} 1 & 0.1109 & -0.1476 \\ & 1 & -0.4199 \\ & & 1 \end{bmatrix}.$$

The coefficients of covariance matrix demonstrate that all the pairs of covariance are in the correct direction and some estimates seem to be underestimated and some are overestimated.

Since the correlations between (X_1, X_2) and (X_1, X_3) were not significantly different from zero, these parameter estimates were reasonable for both models and give a better fit as compared to the Poisson-log normal model (loglikelihood –397.8; Aitchison and Ho, 1989).

7.5.3 Weed species data

The observed covariance matrix and the observed correlation matrix for the weed species data (Chapter 6) are given below:

$$\begin{bmatrix} 2.8099 & 0.0161 & 0.1056 \\ 0.0161 & 0.3481 & -0.1029 \\ 0.1056 & -0.1029 & 27.7325 \end{bmatrix} \qquad \begin{bmatrix} 1 & 0.0162 & 0.0119 \\ & 1 & -0.0331 \\ & & 1 \end{bmatrix}.$$

As mentioned in Table 6.3 (section 6.2), all the pairs of the correlation coefficients were not significantly different from zero. The covariance and the correlation matrices for the different models presented in Chapter 6 are listed below.

(a) Finite mixture with the four components restricted model (AIC selection)

The estimated covariance matrix and the estimated correlation matrix are given below:

$$\begin{bmatrix} 1.9440 & 0.2232 & 0.6911 \\ 0.2232 & 0.3620 & 0.3123 \\ 0.6911 & 0.3123 & 21.6737 \end{bmatrix} \qquad \begin{bmatrix} 1 & 0.2661 & 0.1065 \\ & 1 & 0.1115 \\ & & 1 \end{bmatrix}.$$

(b) Finite mixture with the five components restricted model (BIC selection)

The estimated covariance matrix and the estimated correlation matrix are given below:

$$\begin{bmatrix} 1.7512 & -0.0293 & 0.5754 \\ -0.0293 & 0.7315 & 0.5485 \\ 0.5754 & 0.5485 & 25.2162 \end{bmatrix} \qquad \begin{bmatrix} 1 & -0.0259 & 0.0866 \\ & 1 & 0.1277 \\ & & 1 \end{bmatrix}.$$

(c) Finite mixture with the five components common model (AIC and BIC selection)

The estimated covariance matrix and the estimated correlation matrix are given below:

$$\begin{bmatrix} 2.4377 & 0.0512 & 0.2293 \\ 0.0512 & 0.2669 & -0.0922 \\ 0.2293 & -0.0922 & 25.3294 \end{bmatrix} \qquad \begin{bmatrix} 1 & 0.0635 & 0.0292 \\ & 1 & -0.0355 \\ & & 1 \end{bmatrix}.$$

(d) Finite mixture with the five components independent model (BIC selection)

The estimated covariance matrix and the estimated correlation matrix are given below:

$$\begin{bmatrix} 2.2517 & 0.0750 & 0.1133 \\ 0.0750 & 0.2878 & -0.0662 \\ 0.1133 & -0.0662 & 25.0697 \end{bmatrix} \qquad \begin{bmatrix} 1 & 0.0932 & 0.0151 \\ & 1 & -0.0246 \\ & & 1 \end{bmatrix}.$$

(e) Finite mixture with the six components independent model (AIC selection)

The estimated covariance matrix and the estimated correlation matrix are given below:

$$\begin{bmatrix} 2.4517 & 0.0563 & 0.2162 \\ 0.0563 & 0.3562 & -0.0659 \\ 0.2162 & -0.0659 & 25.3373 \end{bmatrix} \qquad \begin{bmatrix} 1 & 0.0602 & 0.0274 \\ & 1 & -0.0219 \\ & & 1 \end{bmatrix}.$$

From the set of models (a)-(e) for the multivariate Poisson finite mixtures, the model with the five components and the local independence which select by the BIC criterion seem to be the best model compared to all the parameters estimates in the observed and the estimated covariance and correlation matrices.

(f) Hidden Markov model with the five components independent model (AIC and BIC selection)

The estimated covariance matrix and the estimated correlation matrix are given below:

$$\begin{bmatrix} 2.2087 & 0.0629 & 0.0590 \\ 0.0629 & 0.2893 & -0.0476 \\ 0.0590 & -0.0476 & 24.6595 \end{bmatrix} \qquad \begin{bmatrix} 1 & 0.0787 & 0.0080 \\ & 1 & -0.0178 \\ & & 1 \end{bmatrix}.$$

(g) Hidden Markov model with the four components restricted model (AIC and BIC selection)

The estimated covariance matrix and the estimated correlation matrix are given below:

$$\begin{bmatrix} 1.6483 & 0.1311 & 0.7095 \\ 0.1311 & 0.4815 & 1.1935 \\ 0.7095 & 1.1935 & 26.7686 \end{bmatrix} \qquad \begin{bmatrix} 1 & 0.1471 & 0.1068 \\ & 1 & 0.3324 \\ & & 1 \end{bmatrix}.$$

(h) Hidden Markov model with the four components common model (BIC selection)

The estimated covariance matrix and the estimated correlation matrix are given below:

$$\begin{bmatrix} 2.1514 & -0.0550 & 0.0904 \\ -0.0550 & 0.2740 & -0.0053 \\ 0.0904 & -0.0053 & 21.2568 \end{bmatrix} \qquad \begin{bmatrix} 1 & -0.0716 & 0.0134 \\ & 1 & -0.0022 \\ & & 1 \end{bmatrix}.$$

(i) Hidden Markov model with the five components common model (AIC selection)

The estimated covariance matrix and the estimated correlation matrix are given below:

$$\begin{bmatrix} 2.2982 & 0.0821 & 0.1455 \\ 0.0821 & 0.2739 & -0.0267 \\ 0.1455 & -0.0267 & 24.6784 \end{bmatrix} \qquad \begin{bmatrix} 1 & 0.1035 & 0.0193 \\ & 1 & -0.0103 \\ & & 1 \end{bmatrix}.$$

From the set of models (f)-(i) for the multivariate Poisson hidden Markov models, the model with the five components local independence (according to the AIC and the BIC selection) seem to be the best model.

In both sets of models, (a) the multivariate Poisson finite mixture model and (b) the multivariate Poisson hidden Markov model, restricted covariance structure does not seem to be a good indication of the data, even though those models have well separated components (Chapter 6, section 6.5).

All the information of goodness of fit criteria

- Selection of number of components/ states
- Separation of components/states
- Estimated covariance and correlation matrices,

taken into account, the following conclusions can be made for weed count data. The multivariate Poisson hidden Markov model with the independent covariance structure and the five state model was the best representation of data, since this model had a higher entropy criterion value compared to the finite mixture model and the estimated

parameters in the covariance matrices were close to the observed one. This multivariate Poisson hidden Markov model also supports the loglinear analysis results. In addition to that hidden Markov models provide the probability of transition from one state to another.

CHAPTER 8

COMPUTATIONAL EFFICIENCY OF MULTIVARIATE POISSON FINITE MIXTURE MODELS AND MULTIVARIATE POISSON HIDDEN MARKOV MODELS

8.1 Introduction

In this chapter, the computational efficiency of the multivariate Poisson finite mixture models and the multivariate Poissin hidden Markov models is discussed. Since the two sets of models: (a) the multivariate Poisson finite mixture model and (b) the multivariate Poisson hidden Markov model are working well in the setting of finding the unknown number of components or states, it is interested to study about the computational efficiency of the models. Karlis and Xekalaki (1999) discussed the computational efficiency of the finite Poisson mixture models with two components for the maximum likelihood estimation via the EM algorithm.

8.2 Calculation of computer time

Five sets of the multivariate data are simulated with different sample sizes, namely n = 50, 100, 200, 500 and 1000. As we discussed before in Chapter 6, 10 different sets of parameter starting values were randomly selected over the range of data values and the

mixing proportions and the transition probabilities were selected according to uniform random numbers and rescaled them to sum up to 1. First for each set of initial values, the algorithm was run for 100 iterations without any convergence criterion. Then, the model parameters with the largest likelihood were selected for the further analysis. For all five sets, once the suitable initial values have been selected the computer time (CPU time-central processing unit time) was recorded after running the algorithm 200 iterations. The time spent for simulating the samples was not included in this calculation. All the calculations were carried with a PC with a Pentium microprocessor, which has 2 GB of random access memory (RAM). CPU time severely depends on the RAM of the computer. The results of the two models, (a) the multivariate finite mixture model (MFM) and (b) the multivariate hidden Markov model (HMM) were reported for the different components and for the different covariance structures. The CPU times were recorded to the order of 1/100 second.

8.3 Results of computational efficiency

Table 8.1- Table 8.3 and Figure 8.1-Figure 8.3 illustrate the independent, the common, and the restricted covariance structure results, respectively, for the models (a) and (b). It is clear that when the sample size increases, the CPU time (in 1/100 second) also increases exponentially for all models. For small sample sizes (n = 50, 100 and 200) models (a) and (b) have similar CPU times (some cases hidden Markov model take more time) regardless of the number of parameters to be estimated. However, when sample sizes increased (n = 500, 1000) it revealed that the hidden Markov model takes

less time compared to the multivariate finite mixture model even though hidden Markov models have more parameters to be estimated. This is due to the computational procedures involved in these two models.

Therefore, in terms of computational efficiency it can be concluded that for small sample sizes, two models, (a) and (b) have same computational efficiency and for large sample sizes, the multivariate Poisson hidden Markov model is more efficient compared to the multivariate Poisson finite mixture model.

Table 8.1: Independent covariance structure –CPU time (of the order of 1/100 second)

n	k =1(components/states)		k =2 (components/states)		k =3 (components/states)	
	MFM	HMM	MFM	HMM	MFM	HMM
50	22.64	23.72	43.63	45.37	67.61	67.34
100	39.63	43.98	76.58	80.58	115.55	118.20
200	85.17	87.98	164.14	168.22	247.35	260.56
500	216.70	212.46	419.98	416.37	599.51	601.13
1000	470.78	427.33	855.31	843.81	1268.11	1212.84

Table 8.2: Common covariance structure –CPU time (of the order of 1/100 second)

n	k =1 (components/states)		k =2 (components/states)		k =3 (components/states)	
	MFM	HMM	MFM	HMM	MFM	HMM
50	2.81	3.42	5.00	5.56	7.19	7.94
100	5.78	6.53	10.01	10.72	14.25	15.33
200	13.10	13.13	21.87	21.64	30.15	30.64
500	43.30	32.09	66.22	53.23	85.33	75.36
1000	123.11	64.22	165.64	106.90	207.79	153.12

Table 8.3: Restricted covariance structure –CPU time (of the order of 1/100 second)

n	k =1 (components/states)		k =2 (components/states)		k =3 (components/states)	
	MFM	HMM	MFM	HMM	MFM	HMM
50	22.43	23.30	45.38	45.25	65.02	69.52
100	39.91	41.28	80.37	79.11	114.08	117.61
200	85.61	90.84	165.52	168.31	247.29	252.14
500	216.04	210.74	408.54	403.72	593.96	600.56
1000	480.50	421.51	850.26	807.11	1276.52	1207.58

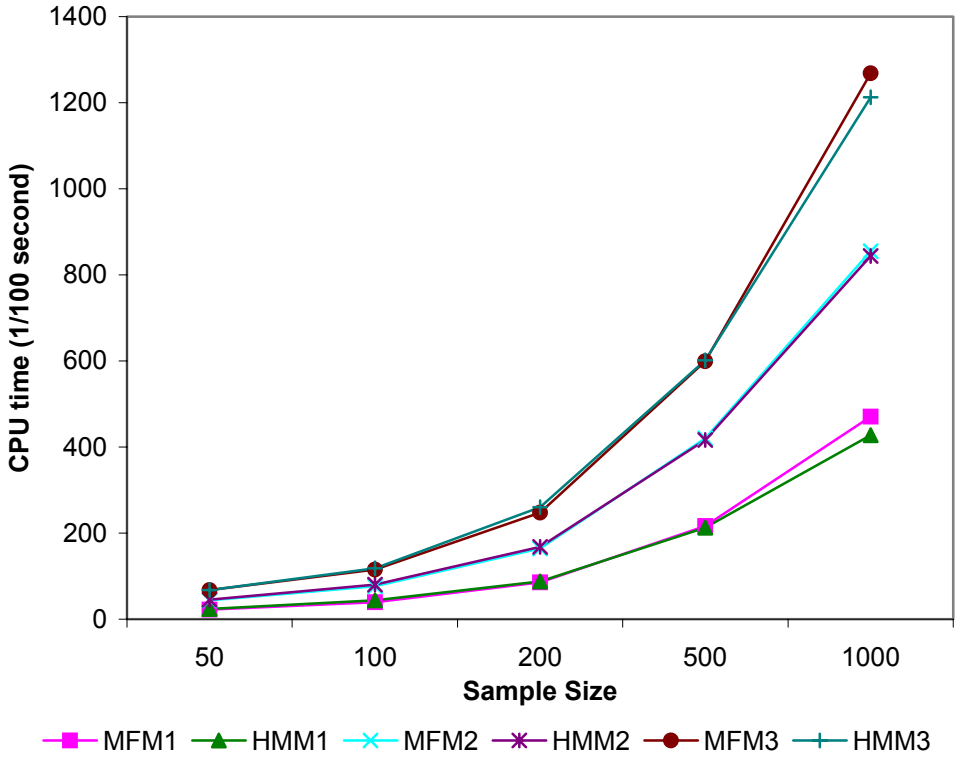


Figure 8.1: Sample Size vs CPU time for the different models of the independent covariance structure

Note:

MFM*n* represents the multivariate Poisson finite mixture model with n components

HMM*n* represents the multivariate Poisson hidden Markov model with n components

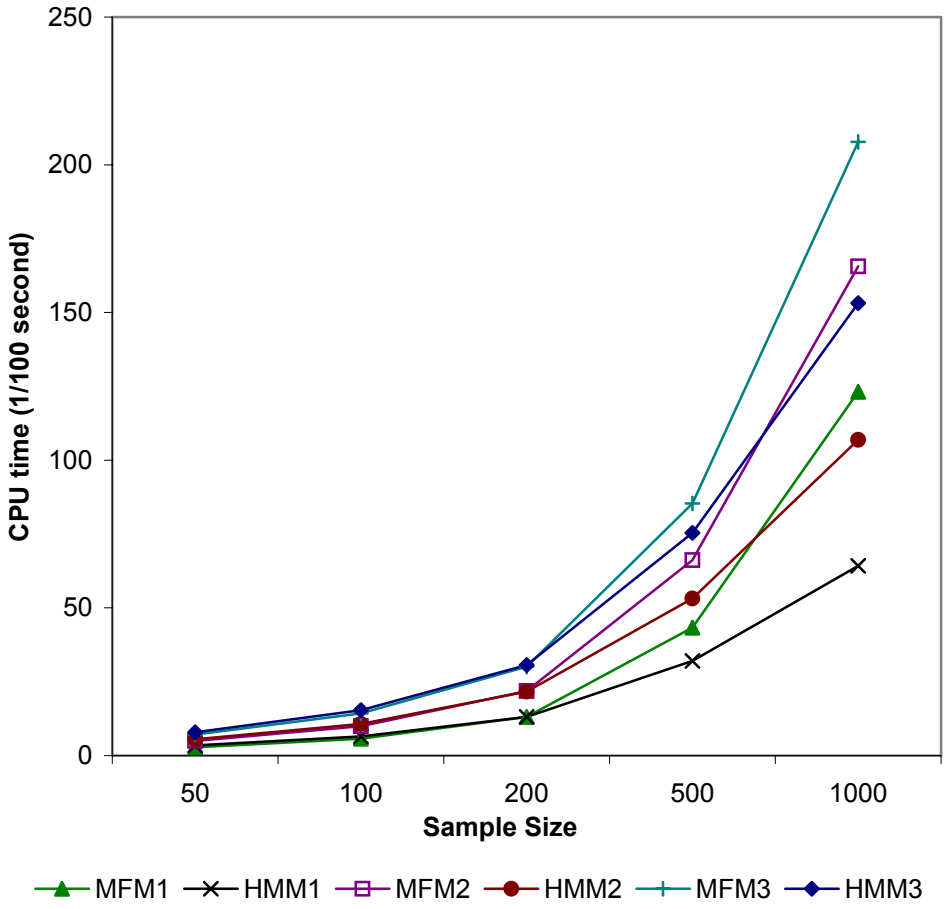


Figure 8.2: Sample Size vs CPU time for the different models of the common covariance structure

Note:

MFMn represents the multivariate Poisson finite mixture model with n components

HMMn represents the multivariate Poisson hidden Markov model with n components

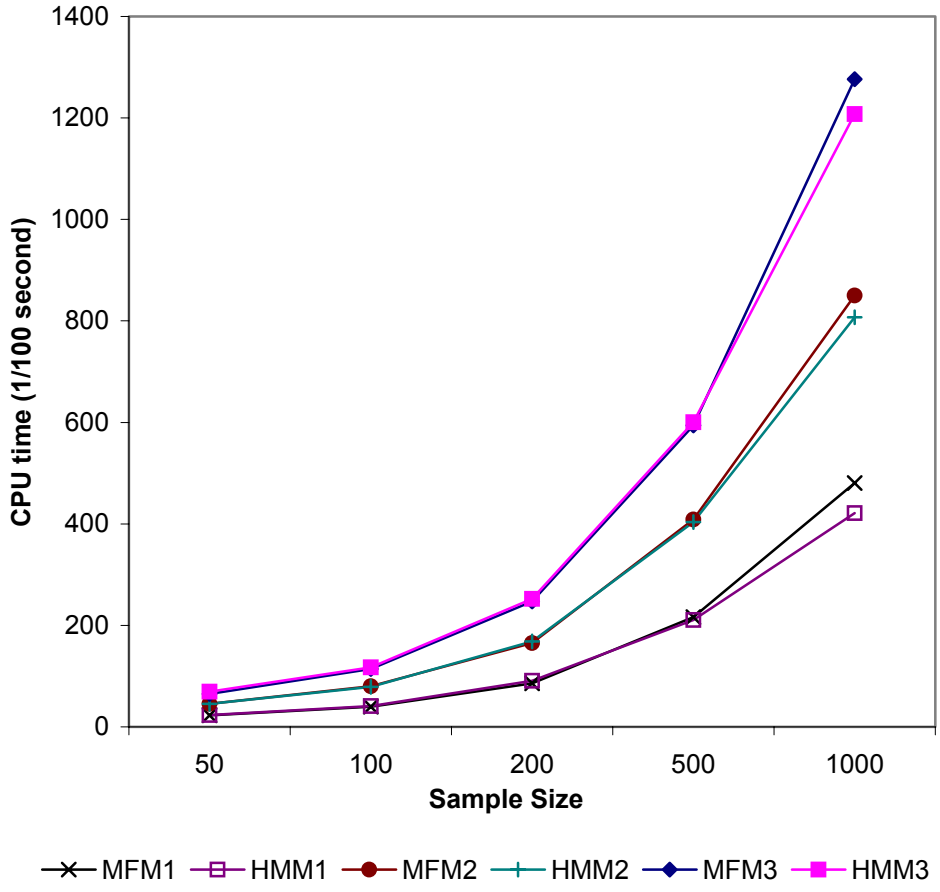


Figure 8.3: Sample Size vs CPU time for different models of restricted covariance structure

Note:

MFMn represents the multivariate Poisson finite mixture model with n components

HMMn represents the multivariate Poisson hidden Markov model with n components

CHAPTER 9

DISCUSSION AND CONCLUSION

9 .1 General summary

Multivariate count data occur in different areas of science. Examples of count data can be found in agriculture (weed species counts in a field), in epidemiology (death count from a disease), in marketing (purchases of different products), in production (different types of faults in a production system), in criminology (different type of crimes in different areas), in accident analysis (different types or different time periods of accidents), and many others. There are a variety of methods available to model the multivariate normal data and the multivariate categorical data. Multivariate count data has small counts with many zeros. Therefore, a normal approximation may not be adequate. The different approaches can be used to handle the multivariate count data. In this study, several more attractive types of models, multivariate Poisson models, were used to overcome the above mentioned problem.

In this thesis, three species counts from an agriculture field were selected for analysis. The main objective was to determine the model for the distribution of these multivariate counts. The estimation involves finding out the mean and covariance structures of the distribution. The data are recorded in a grid, and this data can be considered as a two-dimensional Markov random field. At the same time, an agricultural field has a large neighbourhood system compared to an image. That is, the distance between the neighbouring points or coordinates in an agricultural field is large compared to the distance between the neighbouring points or coordinates in an image. A drawback of the models based on a Markov random field is that they can only be used for small neighbourhoods in an image, due to the computational complexity and the modeling problems posed by large neighbourhoods (Aas et al., 1999). These data can be transformed into a one-dimensional chain. Therefore, as a first step, the grid data were converted into a sequence or a one-dimensional chain using line scan (Chapter 4).

The analysis of these data involves two methods, (a) the multivariate Poisson finite mixture model and (b) the multivariate Poisson hidden Markov model. The multivariate Poisson finite mixture model has been used in many other applications (e.g. marketing Brijs et al., 2004). However, the multivariate Poisson hidden Markov model is a new application to this kind of data (agricultural field data) with Poisson counts.

For both models, the computation of the multivariate Poisson probabilities was studied according to Mahamunulu's recurrence relations (see section 5.2.2). The preliminary loglinear analysis suggests that there were no significant two-way interactions. It can be

seen that p values of the goodness of fit statistic of the models with some two-fold interactions and without any interactions do not differ very much. We decided to include all two-way interactions (section 6.2). Besides, two other interesting covariance structures (common and independent structures which described in section 5.1.2. and 5.1.3) were considered.

9.2 Parameter estimation

Multivariate Poisson hidden Markov model is a special case of the hidden Markov model. The estimation of the parameters of a hidden Markov model most efficiently has done using the likelihood maximization. Baum and Eagon (1967) applied the EM algorithm for locating a local maximum of the likelihood function for a probabilistic function of a Markov chain. Baum et al. (1970) developed the EM algorithm, and applied it to general hidden Markov model. The large-sample behaviour of a sequence of maximum likelihood estimators for a probabilistic function of a Markov chain was studied in Baum and Petrie (1966) and in Petrie (1969). Lindgren (1978) proved a consistency property of maximum likelihood estimators obtained for the model, which assumes that $\{Y_i\}$ is an independent sequence from a finite mixture distribution. Properties of the general ergodic hidden Markov models have been proven: the consistency of the maximum likelihood estimators was proven by Leroux (1992a), and the asymptotic normality of the maximum likelihood estimators was proven by Bickel et al. (1998). Details of the maximum likelihood estimation of the hidden Markov model are found in Leroux (1992 b).

In the applications of the HMMs, likelihood functions and estimates of the model parameters have been routinely computed. However, not much attention has been paid to the computation of standard errors and confidence intervals for parameter estimates of the HMMs (Aittokallio et al., 2000 and Visser et al., 2000). In this thesis, parametric bootstrap samples were generated according to Efron et al., (1993) and McLachlan et al., (2000) and the standard errors of parameter estimates were computed (section 5.5, section 6.3.1 and 6.3.2). These standard errors will be useful for further inferences.

In Chapter 7, we can see that the EM algorithm was performing well for the given dataset (weed counts), for the lens faults dataset (Aitchison and Ho (1989), p.649) and for the bacterial count dataset (Aitchison and Ho (1989), p.651) even though it has some disadvantages (section 5.3.3.1). This analysis also could be done using other optimization techniques such as simulated annealing. However, there is no guarantee that this method is suitable for all kinds of data (Brooks and Morgan, 1995). The EM algorithm has some appealing properties, such as improvement in every iteration and the parameters are in the 'admissible range,' and easy to program (section 5.3.3.1).

9.3 Comparison of different models

There are different ways to handle differences of the fit of the two models. The most well known test is the likelihood ratio test (LRT). Under the null hypothesis (i.e. the fit of both models is equal), the LRT is asymptotically distributed as chi-square with degrees of freedom equal to the difference in the number of parameters if one model is

nested in the other. Since, for instance, local independence model is nested in the common covariance model by deleting the common interaction parameter; this therefore seems like a reasonable test. However regularity conditions needed to use the LRT are not satisfied, because the parameters that allow going from one model to the other take a value at the boundary of the parameter space. Recall that the parameters of any multivariate Poisson model are positive, so the value 0 is at the boundary. This makes the use of the LRT statistic impossible. The same problem arises when testing for the model fit between different component solutions and is well documented in the literature (McLachlan and Peel, 2000).

Another solution for the goodness of fit of the model might be constructing some type of information criterion, like the AIC and the BIC to test the difference between the models. However, these information criterias compare point estimates and not the difference between entire curves, so this does not seem to be applicable either. Therefore, the one way of comparing the different solutions is by visually inspecting loglikelihoods. Figure 6.12 indeed illustrates that the loglikelihood of the independence and the common covariance models clearly dominate the loglikelihoods of the restricted covariance model over the range of component solutions (k=1 to 7). Figure 6.13 illustrates that the loglikelihood of the independence model clearly dominates the loglikelihoods of the restricted and the common covariance model over the range of component solutions (k=1 to 7) for the Markov-dependent models. Viewpoints of the model fit this partially justifies the use of the model with the independent covariance structure since the comparison of maximized loglikelihood providing at least a rough

indication of the relative goodness of fit. The same conclusion was gained after the primarily loglinear analysis and the correlation matrix of the data.

Besides the loglikelihood, when comparing the models, all information about different goodness of fit criterions used in the analysis is listed below:

- Selection of number of components/ states
- Separation of components/states
- Estimated covariance and correlation matrices.

Taking all this information into account, the following conclusion can be made for weed count data. The multivariate Poisson hidden Markov model with the independent covariance structure and the five states is the best representation of the data. This model has the higher entropy index (section 6.5) compared to the finite mixture model and the estimated parameters in the covariance matrix are close to the observed covariance matrix. This model also supports the loglinear analysis results. In addition to that, the hidden Markov model provides the probability of transition from one state to another.

In addition, in terms of the computational efficiency, for the small sample sizes two models, (a) the hidden Markov model and (b) the finite mixture model had similar computational efficiency with respect to the time and for the large sample sizes the hidden Markov model is more computationally efficient compared to the multivariate finite mixture model.

The multivariate Poisson hidden Markov model has some improvements over the existing the multivariate finite mixture model. The computation time of the model is less in the hidden Markov model compared to the finite mixture model for large sample sizes (section 8.3). Another feature of the multivariate Poisson hidden Markov model is that it can take into account the serial correlation among observations and provide the transition probabilities from one state to another.

9.4 Model application to different datasets

The multivariate Poisson finite mixture and the multivariate Poisson hidden Markov models provided a better fit than the multivariate Poisson-log normal model of Aitchison and Ho (1989). The Newton-Raphson method is used to calculate the parameters of the multivariate Poisson-log normal model (section 7.5).

9.5 Real world applications

In general, the multivariate count data occur in different fields of study. In this thesis, we focused on counts for three weed species found in an agricultural field. Even though we selected: Wild Buckwheat, Dandelion and Wild Oats as examples, we can generalize this method to other weed counts as well. The main objective was to find out the distribution of these species. The multivariate Poisson finite mixture models and the multivariate Poisson hidden Markov models are two clustering methods to unmix the distribution and to find parameters and the number of components or states given the

underlying data. The advantage of the hidden Markov model is that it takes serial correlation into account and by introducing suitable covariance structure the idea of the spatial information can be found. Although I have applied this model to "weed counts" it could easily be applied to other datasets. For example, consider an outbreak of a viral infection from a health dataset. A health region is covered by a grid and one can observe the number of cases infecting within a small neighbourhood of each grid point in the health region. The data could also be multivariate if there were several viral infections occurring across the region.

The model suggested in this thesis, the multivariate Poisson hidden Markov model, provides the pattern of the weed distribution. It also gives rates and positive covariance or relationships for weed species within the state. Unconditional covariance matrix for the independent covariance structure shows that there is a negative correlation between Dandelion and Wild Oats. Also, the independent model provides the probability of moving from state i to state j, called transition probabilities. This model could demonstrate how species switch from one component to another, that is, move from one position to another over time.

Our model, the multivariate Poisson hidden Markov model, can deal with both the overdispersion and the spatial information of the data. Therefore this model together with the GIS (geographic information systems) generated weed density maps, will help researchers and farmers to get an insight of weed distributions for herbicide applications. The benefits of this technology include a reduction in spray volume and

consequently lower herbicide costs, timesaving because of fewer stops to refill, less non-target spraying, which reduces potential environmental risks. Further, this model may lead researchers to find other factors, such as soil moisture and fertilizer levels, to determine the states.

The modified EM algorithm for the multivariate Poisson distribution was used to estimate the parameters. There are some problems with the EM algorithm for this model, such as the non-convergence to global optimum and slow convergence. Convergence and the properties of convergence depend heavily on the starting values. Therefore, further studies can be focused on the different optimization techniques for the multivariate Poisson hidden Markov model.

Another disadvantage of the proposed modeling is the lack of availability of software. The modelling of this multivariate Poisson hidden Markov model cannot be done in a user friendly way and one has to write their own code to solve the problem. However, now on, the public can use our Splus/R codes for the analysis of the multivariate Poisson hidden Markov model. Also for the small datasets, this model may not provide the better estimates, since the most of the hidden Markov model properties proved under the assumption of asymptotic behavior.

9.6 Further research

We can present a guideline for the analysis of the multivariate count data (bivariate and trivariate) in different research areas, where the finding the patterns of count data is needed.

Step 1: Exploratory data analysis using histograms, correlations, means and standard deviation of the variables gives the view of the data set you have on your hands.

Step 2: Run the univariate Poisson hidden Markov models for each count variable to see how many mixtures exist.

Step 3: Carry out the loglinear analysis to find out what is the best-described covariance structure for the dataset (restricted, common, or independent).

Step 4: Then fit the multivariate Poisson hidden Markov models for the selected covariance structure.

Step 5: The best-fitted model for the data set can be selected accordingly to the entropy criterion of separation of states and the goodness of fit index of the estimated covariance and the correlation matrix.

As further research, one can map the means and the covariances of the distributions as a layer of a GIS map if the longitude and the latitude coordinates are available. These maps can be applied to the more effective weed control in agricultural fields.

Also the calculations of the multivariate Poisson probabilities of higher dimensions (four or more) can be carefully studied and programmed for further research of the hidden Markov model.

Even though our model provides overall positive and negative correlation of the variables (in this case, species) it is not able to provide the negative interdependence within the components. Therefore, one can study further about this area for Poisson count variables.

The models presented in this thesis do not include any other covariates. The covariates may help us to explain the differences between the distributions of the component. The above issue is high on the list of topics for further research.

REFERENCES

1. Aas, K., Eikvil, L. and Huseby, R. (1999). Applications of hidden Markov chains in image analysis, *Patten Recognition, Vol. 32*(4), pp. 703-713.
2. Agresti, A., (2002). Categorical Data Analysis, New Jersey: John Wiley & Sons, Inc.
3. Aitchison J. and Ho C.H. (1989). The multivariate Poisson-log normal distribution, Biometrika, Vol. 76 (4), pp. 643-653.
4. Aitkin, M. and Aitkin, I. (1996). An hybrid EM/Gauss-Newton algorithm for maximum likelihood in mixture distributions, *Statistics and Computing, Vol. 6*, pp. 127-130.
5. Aittokallio T. and Uusipaikka E. (2000). Computation of Standard Errors for maximum likelihood estimating in Hidden Markov models, Turku Centre for computer Science, Finland, *TUCS Technical Report No. 379*, ISBN 952-12-0760-4.
6. Aiyer A., Pyun K., Huang Y. and Gray D.B.O.R. (2001). Lloyd clustering of gauss mixture models for image compression and classification, IEEE 2001 International conference on Image Processing. [on line]
7. Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B.N. Petrov and F. Csake (eds.), *Second International Symposium on Information Theory. Budapest*: Akademiai Kiado, pp. 267-281.
8. Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, Vol. AC-19*, pp.716-723.

9. Akobundu, I.O., Ekeleme, F. and Chikoye, D. (1999). Influence of fallow management systems and frequency of cropping on weed growth and crop yield. *Weed Research,Vol. 39*, pp.241–256.
10. Archer G.E.B. and Titterington D.M. (2002). Parameter estimation for hidden Markov chains**,** *Journal of Statistical Planning and Inference, Vol. 108*(1), pp.365-390.
11. Banfield, J.D. and Raftery, A.E. (1993). Model-Based Gaussian and Non-Gaussian Clustering, *Biometrics, Vol. 49*, pp. 803-821.
12. Bar-Joseph, Z. and Cohen-Or, D. (2003). Hierarchical Context-based Pixel Ordering, *Computer Graphics Forum, Proceedings of Eurographics, Vol. 22*(3), pp. 1-10.
13. Basford, K.E., Greenway, D.R., McLachlan, G.J. and Peel, D. (1997a). Standard errors of fitted means under normal mixture models, *Computational Statistics, Vol. 12*, pp. 1-17.
14. Baum, L.E. and Petrie, T. (1966). Statistical inference for probabilistic functions of finite state Markov chains, *Annals of Mathematical Statistics, Vol. 37*, pp.1554-1563.
15. Baum, L.E., Peterie, T., Souled, G. and Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains, *Annals of Mathematical Statistics*, *Vol. 41*(1), pp. 164-171.
16. Baum, L.E. and Egon, J.A. (1967). An inequality with applications to statistical estimation for probabilistic functions of a Markov process and to a model for ecology, *Bulletin of the American Meteorological Society*, *Vol.73*, pp. 360-363.

17. Bicego, M., Murino, V. & Figueiredo, M.A.T (2003). A sequential pruning strategy for the selection of the number of states in hidden Markov models**,** *Pattern Recognition Letters, Vol. 24*(9), pp. 1395-1407.
18. Bickel P.J., Ritov Y., and Rydén T. (1998). Asymptotic Normality of the maximum likelihood estimator for general hidden Markov models, *The Annals of Statistics, Vol. 26* (4), pp. 1614-1635.
19. Binder, K. (1979). Monte Carlo Methods in Statistical Physics, New York: Springer-Verlag.
20. Binder, K. and Heermann, D.W. (1992). Monte Carlo Simulation in Statistical Physics, New York: Springer-Verlag.
21. Brijs, T. (2002). Retail Market Basket Analysis: A Quantitative Modelling Approach, *Ph.D. dissertation*, Faculty of Applied Economics, Limburg University Center, Belgium.
22. Brijs, T., Karlis, D., Swinnen, G., Vanhoof, K., Wets, G., and Manchanda, P. (2004). A multivariate Poisson mixture for marketing applications, *Statistica Neerlandica*. *Vol. 58*(3), pp. 322-348.
23. Brooks S.P. and Morgan B.J.T. (1995). Optimization Using Simulated Annealing, *The Statistician, Vol. 44* (2), pp. 241-257.
24. Campbell, J.G., Fraley, C., Murtagh, F., and Raftery, A.E. (1997). Linear Flaw Detection in Woven Textiles using Model-Based Clustering, *Pattern Recognition Letters, Vol. 18*, pp.1539-1548.

25. Cappé, O., (2001). H2M: A set of MATLAB/OCTAVE functions for the EM estimation of mixtures and hidden Markov Models, *Technical Report, ENST Department*, TSI/LTCI, Paris, France.
26. Celeux, G. and Govaert, G. (1995). Gaussian parsimonious clustering models. *Pattern Recognition, Vol. 28*, pp.781-793.
27. Chernick, M.R. (1999). Bootstrap Methods: A Practitioner's Guide, New York: John Wiley & Sons.
28. Chib S. and Winkelmann R. (2001). Markov Chain Monte Carlo Analysis of Correlated Count Data, Journal of Business and Economic Statistics, Vol. 19 (4), pp. 428-435.
29. Dasgupta, A. and Raftery, A.E. (1998). Detecting Features in Spatial Point Processes with Clutter via Model-Based Clustering. *Journal of the American Statistical Association, Vol. 93*(441), pp. 294-302.
30. Davison, A.C. and Hinkley, D.V. (1997). Bootstrap methods and their applications, Cambridge, UK; New York, NY, USA: Cambridge University Press.
31. de Rouw, A**.** (1995). The fallow period as a weed-break in shifting cultivation (tropical wet forests). *Agriculture, Ecosystems, and Environment, Vol. 54*, pp. 31–43.
32. Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm*, Journal of the Royal Statistical Society. Series B (Methodological)*, *Vol. 39*(1), pp. 1-38.

33. Descombes, X., Morris, R.D., Zerubia, J., and Berthod, M. (1999). Estimation of Markov Random Field Prior Parameters using Markov Chain Monte Carlo Maximum likelihood, IEEE Transactions of Image Processing, Vol. 8(7), pp.954-963.

34. Diobolt, J. and Ip, E.H.S. (1996). Stochastic EM: Method and Application, *Markov Chain Monte Carlo in Practice*, London: Chapman and Hall, pp.259-274.

35. Efron, B. and Tibshirani, R. (1993). An Introduction to the Bootstrap, London: Chapman and Hall.

36. Efron, B. (1979). Bootstrap methods:another look at the jackknife, *Annals of Statistics, Vol. 7*, pp.1-26.

37. Efron, B. (1982). The Jackknife, the Bootstrap and other Resampling Plans, Philadelphia: Series#:38; Cbms-Nsf Regional Conference Series in Applied Mathematics; SIAM Society for Industrial & Applied Mathematics (SIAM).

38. Elliott, R.J., Aggoun L., and Moore J.B. (1995). Hidden Markov Models: Estimation and Control, New York: Springer-Verlag.

39. Elliott, R.J. and Aggoun L. (1996). Estimation for Hidden Markov Random Fields, *Journal of Statistical Planning and Inference*, *Vol. 50*, pp. 343-351.

40. Engel, C. and Hamilton, J.D. (1990). Long Swings in the Dollar: Are They in the Data and Do Markets Know it? , *American Economic Review, Vol. 80*, pp. 689-713.

41. Felsenstein, J. and Churchill, G.A. (1996). A hidden Markov Model approach to variation among sites in rate of evolution, *Molecular Biology and Evolution, Vol. 13*, pp. 93-104.

42. Fishman, G.S., (1996). Monte Carlo: Concepts, algorithms, and applications, New York: Springer-Verlag.

43. Fjϕrtoft, R., Boucher, J., Delignon, Y., Garello, R., Le Caillec, J., Maitre, H., Nicolas, J., Pieczynski, W., Sigelle, M., and Tupin, F. (2000). Unsupervised Classification of Radar Images Based on Hidden Markov Models and Generalized Mixture Estimation*, Proc. EOS/SPIE Symposium on Remote Sensing, Conference on SAR Image Analysis, Modelling, and Techniques V*, *Vol. SPIE 4173*, pp. 87-98.

44. Fjϕrtoft, R., Delignon, Y., Pieczynski, W., Sigelle, M. and Tupin, F. (2003). Unsupervised Classification of Radar Images Using Hidden Markov Chains and Hidden Markov Fields, *IEEE Transactions on Geoscience and Remote Sensing, Vol. 41*(3), pp. 675-686.

45. Fjϕrtoft, R., Wojciech Pieczynski and Yves Delignon (2001), Generalised Mixture Estimation and Unsupervised Classification Based on Hidden Markov Chains and Hidden Markov Random Fields, Proc. *Scandinavian Conference on Image Analysis (SCIA'01)*, pp. 733-740

46. Forney, Jr. G.D. (1973). The Viterbi Algorithm, *Proceedings of IEEE, Vol. 61*(3), pp. 263-278.

47. Fraley, C. and Raftery, A.E. (1998). How many clusters? Which clustering method? Answers via Model based Cluster Analysis, *Technical Report No. 329,*

Department of Statistics, University of Washington, Box 354322, Seattle, WA, 98193-4322, USA.

48. Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *Vol. 6*(6), pp.721-741.
49. Hasselblad, V., (1969). Estimation of Finite Mixtures of Distributions from the Exponential Family, *Journal of the American Statistical Association, Vol. 64*(328), pp. 1459-1471.
50. Holgate P. (1964). Estimation for the bivariate Poisson distribution, *Biometrika, Vol. 51* (1), pp.241-245.
51. Johnson N.L. and Kotz S. (1969). Distributions in Statistics: Discrete Distributions. Boston: Houghton Mifflin.
52. Johnson, N.L., Kemp, A.W. and Kotz, S. (2005). *Univariate Discrete Distributions*, Hoboken, New Jercy: Wiley-Interscience.
53. Johnson, N.L., Kotz, S. and Balakrisnan, N. (1997). Discrete Multivariate Distributions, New York: John Wiley and Sons, pp. 1-30, 124-152.
54. Juang, B.H. and Rabiner, L.R.(1991). Hidden Markov Models for Speech Recognition, *Technometrics, American Statistical Association and the American Society for Quality Control, Vol. 33*(3), pp. 251-272.
55. Kano, K. and Kawamura, K. (1991). On Recurrence Relations for the Probability Function of Multivariate Generalized Poisson distribution, *Communications in Statistics: Theory and Methods, Vol. 20*(1), pp. 165-178.

56. Karlis D. (2003). An EM Algorithm for Multivariate Poisson Distribution and Related Models, *Journal of Applied Statistics, Vol. 30,* pp. 63-77.

57. Karlis D. and Meligkotsidou L. (2005). Multivariate Poisson regeression with covariance structure, *Statistics and Computing, Vol. 15*, pp. 225-265.

58. Karlis D. and Xekalaki E. (1998). An improvement of the EM algorithm for finite Poisson mixtures, *Proceeding of HERCMA 98, Athens,* pp. 596-604.

59. Karlis D., and Xekalaki E., (1999). Improving the EM algorithm for mixtures, *Statistics and Computing, Vol. 9***,** pp.303-307.

60. Karlis, D. and Meligkotsidou, L. (2006). Finite Mixtures of Multivariate Poisson Distributions with Application, *Journal of Statistical Planning and Inference*, doi:10.1016/j.jspi.2006.07.001.

61. Kunsch, H., Geman, S., and Kehagias, A. (1995). Hidden Markov Random Fields, *The Annals of applied Probability, Vol. 5*(3), pp. 577-602.

62. Laird, N.M. (1978). Nonparametric Maximum Likelihood estimation of a mixing distribution, *Journal of the American Statistical Association, Vol. 73*(364), pp.805-811.

63. Leroux, B.G. (1992a). Consistent Estimation of a Mixing Distribution, *The Annals of Statistics, Vol. 20*(3), pp. 1350-1360.

64. Leroux, B.G. (1992b). Maximum- likelihood Estimation of Hidden Markov Models, *Stochastic Processes and their Applications, Vol. 40*, pp. 127-143.

65. Leroux, B.G. and Puterman, M. L. (1992). Maximum Penalized likelihood Estimation for Independent and Markov-Dependent Mixture models, *Biometrics, Vol. 48*, pp. 545-558.

66. Li C.S., Lu J.C., Park J., Kim K., Brinkley P.A. and Peterson J.P. (1999). Multivariate zero-inflated Poisson models and their applications, *American Statistical Association, Vol. 41* (1), pp.29-38.

67. Lindgren G. (1978). Markov regime models for mixed distributions and switching regressions, *Scandinavian Journal of Statistics, Vol. 5*, pp. 81-91.

68. Mackay, R.J., (2002). Estimating the order of a hidden Markov Model, *The Canadian Journal of Statistics, Vol. 30*, pp. 573-589.

69. Mahamunula, D.M. (1967). A Note on Regression in the Multivariate Poisson Distribution, *Journal of the American Statistical Association*, *Vol. 62*(317), pp. 251-258.

70. Mardia, K.V. (1970). Families of Bivariate Distributions, London: Charles Griffin & Compant Limited.

71. Marshall, A.W., and Olkin, I. (1985). A family of bivariate distributions generated by the bivariate Bernouilli distributions, *Journal of the American Statistical Association, Vol. 80*(390), pp. 332-338.

72. McCoy, B.M. and Wu, T.T. (1973). The Two-Dimentional Ising Model, Harvard University Press, Cambridge Massachusetts.

73. McHugh, R.B. (1956). Efficient estimation and local identification in latent class analysis, *Psychometrika, Vol. 21*, pp. 331-347.

74. McLachlan, G. (1982). The Classification and Mixture Maximum Likelihood Approaches to Cluster Ananlysis, Handbook of Statistics, Vol. 2, pp. 199-208.

75. McLachlan, G. and Basford, K.E., (1988). Mixture Models: Inference and Applications to Clustering, New York: John Wiley & Sons, Inc.

76. McLachlan, G. and Krishnan, T. (1997). The EM Algorithm and Extensions, New York: John Wiley and Sons.

77. McLachlan, G. and Peel, D. (2000). Finite Mixture Models, New York: John Wiley & Sons, Inc.

78. Meilijson, I. (1989). A fast improvement of the EM on its own terms, *Journal of the Royal Statistical Society, B51*, pp.127-138.

79. Memon, N., Neuhoff, D.L. and Shende, S. (2000). An analysis of some common scanning techniques for lossless image coding, *IEEE Transactions on Image Processing, Vol. 9*(11), pp. 1837-1848.

80. Murtagh, F. and Raftery, A.E. (1984). Fitting straight lines to the point patterns. *Pattern Recognition, Vol. 17*, pp.479-483.

81. Ngobo, M., McDonald, M. and Weise, S. (2004). Impacts of type of fallow and invasion by Chromolaena odorata on weed communities in crop fields in Cameroon, Ecology and Society, Vol. **9**(2):[online] Available from: http://www.ecologyandsociety.org/vol9/iss2/art1/

82. Pearson, K. (1894). Contributions to the mathematical theory of evolution, Philosophical Transactions, *Vol. A185*, pp. 71-110.

83. Peel, D. (1998). Mixture Model Clustering and Related Topics, *Unpulished Ph.D. Thesis, University of Queensland*, Brisbane.

84. Permuter H., Francos J., and Jermyn I. (2006). A study of Gaussian mixture models of color and texture features for image classification and segmentation, *Pattern Recognition, Vol. 39*, pp. 695-706.

85. Petrie T. (1969). Probabilistic functions of finite state Markov chains, Annals of Mathematical Statistics, Vol. 40, pp.97-115.

86. Petrushin, V.A., (2000). Hidden Markov Models: Fundamentals and Applications- Part 1: markov chains and Mixture Models, Online Symposium for Electronics Engineer [online],

Available from: http://www.techonline.com/asee/ [accessed April 25 2005].

87. Pieczynski W., Benboudjema D., and Lanchantin P. (2002). Statistical image segmentation using Triplet Markov fields, SPIE's International Symposium on Remote Sensing, September 22-27, Crete, Greece.

88. Rabiner, L.R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition, *Proceedings of the IEEE, Vol. 77*(2), pp.257-286.

89. Rabiner, L.R. and Juang, B.H., (1986). An Introduction to Hidden Markov Models, *IEEE ASSP Magazine, Vol. 3*(1), pp 4-16.

90. Reynolds D.A., Quatieri T.F. and Dunn R.B. (2000). Speaker verification using adapted Gaussian mixture models, *Digital Signal Processing, Vol. 10*, pp.19-41.

91. Robert, C.P. (1996). Mixture of distributions: Inference and Estimation, *Markov Chain Monte Carlo in Practice*, London: Chapman and Hall, pp.441-464.

92. Ross, S. M. (1996). Stochastic Processes, New York: Wiley & Sons, Inc.

93. SAS/STAT software (2003). Statistical Analysis System, Release 8.2, SAS Institute Inc. Cary, NC, USA.

94. Saul L.K. and Lee D.D. (2001). Multiplicative updates for classification by mixture models, *Advances in Neural Information Processing Systems 14, Vol. 2*, pp. 897-904.

95. Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, Vol. 6*, pp. 461-464.

96. Seidel, W., Mosler, K. and Alker, M. (2000). A cautionary note on likelihood ratio tests in mixture models, *Annals of the Institute of Statistical Mathematics, Vol. 52*, pp.481-487.

97. Srinivasan, S.K. and Mehata, K.M. (1978). Stochastic Processes, New York: McGraw-Hill.

98. Stokes, M.E., Davis, C.S., and Koch, G.G. (2000). Categorical Data Analysis Using the SAS System, 2nd edition, SAS Institute, SAS publishing.

99. Symons, M.J. (1981). Clustering Criteria and Multivariate Normal Mixtures, Biometrics, Vol. 37, pp. 35-47.

100. Symons, M.J., Grimson, R.C. and Yuan, Y.C. (1983). Clustering of rare events, *Biometrics, Vol. 39*, pp. 193-205.

101. Titterington, D.M., Smith, A.F.M., and Markov, U.E., (1985). Statistical Analysis of Finite Mixture Distributions, Great Britain: John Wiley & Sons.

102. Titterington, D.M. (1990). Some recent research in the analysis of mixture distributions, *Statistics*, *Vol. 4*, pp. 619-641.

103. Tsiamyrtzis, P. and Karlis, D. (2004). Strategies for efficient computation of multivariate Poisson probabilities, *Communications in Statistics, Simulation and Computation, Vol. 33,* pp.271 –293.

104. University of Manitoba, Department of Plant Science (2005). Weed Identification Library. Retrieved March 5, 2005, from http:// www.umanitoba.ca/afs/plant_science/weeds/ weeds.html.

105. Vermunt, J.K. and Magidson, J. (2002). Latent class cluster analysis, Chapter 3, in Hagenaars, J.A., and McCutche, A.L. (editors) on *Applied Latent Class Anlysis*, Cambridge University Press.

106. Visser I., Raijmakers M.E.J. and Molenaar P.C.M. (2000). Confidence Intervals for Hidden Markov Model Parameters, *British Journal of Mathematical and Statistical Psychology, Vol. 53*, pp. 317-327.

107. Viterbi, A.J. (1967). Error bounds for convolution codes and an asymptotically optimal decoding algorithm, *IEEE Trans. Info. Theory, Vol. 13*, pp.260-269.

108. Wu, C.F.J. (1983). On the Convergence Properties of the EM Algorithm, *The Annals of Statistics, Vol. 11*(1), pp. 95-103.

109. Wu, F.Y. (1982). The Potts Model, *Reviews of Modern Physics*, Vol. 54, pp. 235-268.

110. Zhang, Y., Brady, M., and Smith, S., (2001). Segmentation of Brain MR Images Through a Hidden Markov Random Field Model and the Expectation-Maximization Algorithm, *IEEE Transactions on Medical Imaging*, *Vol. 20*(1), pp.45-57.

111. Zucchini W., Adamson, P. & McNeill, L. (1991). A Family of Models for Drought, *Journal of Plants and Soil, Vol. 27*, pp. 1917-1923.

APPENDIX

A. Splus/R code for Multivaraite Poisson Hidden Markov Model- Common Covariance Structure (This is an example for three components)

###Markov dependent+ 3 components+ common covariance###

##**NOTE: Before you start implementing the code please read the Chapter 5**##

data<-read.table("j://data1.txt",header=T) # Read the data from the text file
attach(data)
y1<-data[,1]
y2<-data[,2]
y3<-data[,3]

######Function to calculate trivariate Poisson probabilities for common covariance
##structure
##[**Refer to Section 5.1.2, Section 5.2.1 and equation (5.4)]**

$p(\mathbf{y};\boldsymbol{\theta}) = P[Y_1 = y_1, Y_2 = y_2, ..., Y_n = y_n]$

$$\# \quad = \exp\left(-\sum_{i=1}^{n}\theta_i\right)\prod_{i=1}^{n}\frac{\theta_i^{y_i}}{y_i!}\sum_{i=0}^{s}\left[\prod_{j=1}^{n}\binom{y_j}{i} i!\left(\frac{\theta_0}{\prod_{k=1}^{n}\theta_k}\right)^{i}\right], \tag{5.4}$$

#where $s = \min\{y_1, y_2, ..., y_n\}$.

ptrivpois<-function(x, y, z, lambda = c(1, 1, 1, 1), log=FALSE) {

EM algorithms for trivariate Poisson Models
----
x : 1st count variable
y : 2nd count variable
z : 3nd count variable
lambda : parameters of the trivariate poisson distribution
log : argument controlling the calculation of the log-probability or the
probability function.
----
n <- length(x)

x0<-x[1]
y0<-y[1]

```
        z0<-z[1]
        xyzmin<-min( x0,y0,z0 )
        lambdaratio<-lambda[4]/(lambda[1]*lambda[2]*lambda[3])

        i<-0:xyzmin
        sums<- -lgamma(x-i+1)-2*lgamma(i+1)-lgamma(y-i+1)-lgamma(z-
i+1)+i*log(lambdaratio)
        maxsums <- max(sums)
        sums<- sums - maxsums
        logsummation<- log( sum(exp(sums)) ) + maxsums
        logtp<- -sum(lambda) + x * log( lambda[1] ) + y * log( lambda[2] )+ z * log(
lambda[3] ) + logsummation
        logtp
        if (log) { result<-    logtp }
        else     { result<-exp(logtp)  }
        result
#       end of function trivpois
}

##-----------------------------------

theta11 <-1.28 # initialize parameter values for component 1
theta21 <-0.25
theta31 <-3
theta41<-0.01

theta12 <-0.5 # initialize parameter values for component 1
theta22 <-0.15
theta32 <-3
theta42<-1

theta13 <-0.15 # initialize parameter values for component 1
theta23 <-1
theta33 <-2.5
theta43<-0.1

##constants##
N<-3      # number of components
T<-150   # number of observations
Nit<-100  # number of  Iterations


#Initial Transition matrix##
TRANS<-matrix(c(0.25,0.5,0.25,0.1,0.5,0.4,0.3,0.2,0.5),nrow=N,ncol=N,byrow=T)

loglike<-rep(0,Nit)
```

#Main loop of the EM algorithm###

for (nit in 1:Nit){

#initialize matrices to store probabilities in different stages
result1<-matrix(NA,nrow=T,ncol=1)
d1231<-matrix(NA,nrow=T,ncol=1)
d1231old<-matrix(NA,nrow=T,ncol=1)
x1231<-matrix(NA,nrow=T,ncol=1)
d1232<-matrix(NA,nrow=T,ncol=1)
d1232old<-matrix(NA,nrow=T,ncol=1)
x1232<-matrix(NA,nrow=T,ncol=1)
d1233<-matrix(NA,nrow=T,ncol=1)
d1233old<-matrix(NA,nrow=T,ncol=1)
x1233<-matrix(NA,nrow=T,ncol=1)

#initialize matrices to store probabilities from three components
threep1<-matrix(NA,nrow=T,ncol=1)
threep2<-matrix(NA,nrow=T,ncol=1)
threep3<-matrix(NA,nrow=T,ncol=1)
py<-matrix(NA,nrow=T,ncol=1) # Store final probability function here
w1<-matrix(NA,nrow=T,ncol=1) #Store posterior probabilities for component 1
w2<-matrix(NA,nrow=T,ncol=1) # Store posterior probabilities for component 2
w3<-matrix(NA,nrow=T,ncol=1) #Store posterior probabilities for component 3

#Initilize matrices to store values of X's
x11<-matrix(NA,nrow=T,ncol=1)
x21<-matrix(NA,nrow=T,ncol=1)
x31<-matrix(NA,nrow=T,ncol=1)
x141<-matrix(NA,nrow=T,ncol=1)
x12<-matrix(NA,nrow=T,ncol=1)
x22<-matrix(NA,nrow=T,ncol=1)
x32<-matrix(NA,nrow=T,ncol=1)
x142<-matrix(NA,nrow=T,ncol=1)
x13<-matrix(NA,nrow=T,ncol=1)
x23<-matrix(NA,nrow=T,ncol=1)
x33<-matrix(NA,nrow=T,ncol=1)
x143<-matrix(NA,nrow=T,ncol=1)

Start the EM algorithm
for (i in 1:T){
threep1[i]<-
(ptrivpois(y1[i],y2[i],y3[i],lambda=c(theta11,theta21,theta31,theta41),log=FALSE))

```
threep2[i]<-
(ptrivpois(y1[i],y2[i],y3[i],lambda=c(theta12,theta22,theta32,theta42),log=FALSE))
threep3[i]<-
(ptrivpois(y1[i],y2[i],y3[i],lambda=c(theta13,theta23,theta33,theta43),log=FALSE))

d1231[i]<-0
d1232[i]<-0
d1233[i]<-0
result1[i]<-min(y1[i],y2[i],y3[i])
for (r in 0:result1[i]){
d1231old[i]<-r*dpois(y1[i]-r,theta11)*dpois(y2[i]-r,theta21)*dpois(y3[i]-
r,theta31)*dpois(r,theta41)
d1231[i]<-d1231old[i]+d1231[i]
d1232old[i]<-r*dpois(y1[i]-r,theta12)*dpois(y2[i]-r,theta22)*dpois(y3[i]-
r,theta32)*dpois(r,theta42)
d1232[i]<-d1232old[i]+d1232[i]
d1233old[i]<-r*dpois(y1[i]-r,theta13)*dpois(y2[i]-r,theta23)*dpois(y3[i]-
r,theta33)*dpois(r,theta43)
d1233[i]<-d1233old[i]+d1233[i]}
```

check the condition for Poisson random variables [**Refer to equation (5.2)**]

$$\begin{aligned} y_1 - x_{12} - x_{13} - x_{123} &\geq 0 \\ y_2 - x_{12} - x_{23} - x_{123} &\geq 0 \\ y_3 - x_{13} - x_{23} - x_{123} &\geq 0. \end{aligned} \qquad (5.2)$$

```
if (threep1[i]==0) {x141[i]<-0} else {x141[i]<-d1231[i]/threep1[i]}
if (threep2[i]==0) {x142[i]<-0} else {x142[i]<-d1232[i]/threep2[i]}
if (threep3[i]==0) {x143[i]<-0} else {x143[i]<-d1233[i]/threep3[i]}


if ((y1[i]-x141[i])>0) {x11[i]<-(y1[i]-x141[i])} else {x11[i]<-y1[i]}

if ((y2[i]-x141[i])>0) {x21[i]<-(y2[i]-x141[i])} else {x21[i]<-y2[i]}

if ((y3[i]-x141[i])>0) {x31[i]<-(y3[i]-x141[i])} else {x31[i]<-y3[i]}

if ((y1[i]-x142[i])>0) {x12[i]<-(y1[i]-x142[i])} else {x12[i]<-y1[i]}

if ((y2[i]-x142[i])>0) {x22[i]<-(y2[i]-x142[i])} else {x22[i]<-y2[i]}

if ((y3[i]-x142[i])>0) {x32[i]<-(y3[i]-x142[i])} else {x32[i]<-y3[i]}

if ((y1[i]-x143[i])>0) {x13[i]<-(y1[i]-x143[i])} else {x13[i]<-y1[i]}
```

if ((y2[i]-x143[i])>0) {x23[i]<-(y2[i]-x143[i])} else {x23[i]<-y2[i]}

if ((y3[i]-x143[i])>0) {x33[i]<-(y3[i]-x143[i])} else {x33[i]<-y3[i]}}

#Initilize forward and backward variables and place to store loglikelihood
logl=matrix(0,nrow=1,ncol=Nit)
dens=matrix(0,nrow=T,ncol=N)
alpha=matrix(0,nrow=T,ncol=N)
beta=matrix(0,nrow=T,ncol=N)
scale=matrix(c(1,rep(0,T-1)),nrow=1,ncol=T)
ones=matrix(1,nrow=T,ncol=1)

####E-step, compute density values
dens=cbind(threep1,threep2,threep3)

####E-step, forward recursion and likelihood computation
##Use a uniform a priori probability for the initial state
#[**Refer to equation (5.25) and (5.26)**]
#The forward and backward variables

$\# \alpha_j(i) = P[\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_n, S_i = j]$ and

$$\# \beta_j(i) = P[\mathbf{y}_{i+1}, ..., \mathbf{y}_n \mid S_i = j] \tag{5.25}$$

#which yield the quantities of interest by

$$\# \hat{u}_j(i) = \frac{\alpha_j(i)\beta_j(i)}{\sum_l \alpha_l(n)} = \frac{\alpha_j(i)\beta_j(i)}{\sum_{j=1}^{m} \alpha_j(i)\beta_j(i)} \quad \text{and}$$

$$\# \hat{v}_{jk}(i) = \frac{P_{jk} f(\mathbf{y}_i; \boldsymbol{\lambda}_k)\alpha_j(i-1)\beta_k(i)}{\sum_l \alpha_l(n)}. \tag{5.26}$$

alpha[1,]=dens[1,]/N
for (t in 2:T){
alpha[t,]=(alpha[t-1,]%%TRANS)*dens[t,]*
#####Systematic scaling
scale[,t]=sum(alpha[t,])
alpha[t,]=alpha[t,]/scale[,t]
}

###compute likelihood

loglike[nit]<-sum(log(scale))

####E-step, backward recursion
###Scale the backward variable with the forward scale factors
###(this ensures that the reestmation of the transition matrix below is correct)
#[**Refer to equation (5.25) and (5.26)**]

beta[T,]=matrix(1,nrow=1,ncol=N)
for (t in (T-1):1){
*beta[t,]=(beta[t+1,]*dens[t+1,])%*%t(TRANS)*
beta[t,]=beta[t,]/scale[,t]
}

####M-step, reestmation of the transition matrix
##compute unnormalized transition probabilities (this is indeed still the end of the E-
#step, which explains that TRANS appears on the right-hand side below)
#[**Refer to equation (5.24)**]

$$\# P_{jk} = \frac{\sum_{i=2}^{n} \hat{v}_{jk}(i)}{\sum_{i=2}^{n}\sum_{l=1}^{m} \hat{v}_{jl}(i)} \ . \qquad (5.24)$$

TRANS=TRANS(t(alpha[1:(T-1),])%*%(dens[2:T,]*beta[2:T,]))*

###Normalization of the transition matrix

oness=matrix(1,nrow=1,ncol=N)
sumtrans=matrix(0,nrow=N,ncol=1,byrow=T)
for (n in 1:N){
sumtrans[n,]=sum(TRANS[n,])
}
sumtrans=sumtrans%%oness*
TRANS=TRANS/sumtrans

####M-step, reestimation of the rates
###Compute a posteriori probabilities and store them in matrix beta to save space
#[**Refer to equation (5.31) and (5.32)**]
#Then M-step computes the posteriori probabilities using the following equation.

$$\#\widehat{u}_j(i) = P[S_i = j \mid \mathbf{Y} = \mathbf{y}] = \frac{\alpha_j(i)\beta_j(i)}{\sum_{l=1}^{m}\alpha_l(n)} = \frac{\alpha_j(i)\beta_j(i)}{\sum_{j=1}^{m}\alpha_j(i)\beta_j(i)} \tag{5.31}$$

#and then re-estimate the rates as follows:

$$\#\hat{\boldsymbol{\lambda}}_j = \frac{\sum_{i=1}^{n}\widehat{u}_j(i)\mathbf{d}_i^j}{\sum_{i=1}^{n}\widehat{u}_j(i)}, \qquad j = 1,...,m. \tag{5.32}$$

```
beta=alpha*beta
sumbeta=matrix(0,nrow=T,ncol=1,byrow=T)
for (r in 1:T){
sumbeta[r,]=sum(beta[r,])
}
sumbeta=sumbeta%*%oness
beta=beta/sumbeta

##Reestmate rates
#component 1
newdata1=cbind(x11,x21,x31,x141)
rate1=(t(beta[,1])%*%newdata1)/sum(beta[,1])
#component 2
newdata2=cbind(x12,x22,x32,x142)
rate2=(t(beta[,2])%*%newdata2)/sum(beta[,2])
#component 3
newdata3=cbind(x13,x23,x33,x143)
rate3=(t(beta[,3])%*%newdata3)/sum(beta[,3])


#Assign estimated parameters new variables
theta11 <-rate1[,1] # component 1 estimates
theta21 <-rate1[,2]
theta31 <-rate1[,3]
theta141<-rate1[,4]

theta12 <-rate2[,1] # component 2 estimates
theta22 <-rate2[,2]
theta32 <-rate2[,3]
theta142<-rate2[,4]

theta13 <-rate3[,1] # component 3 estimates
theta23 <-rate3[,2]
```

theta33 <-rate3[,3]
theta143<-rate3[,4]
}
####RESULTS####
loglike
rate1
rate2
rate3
TRANS

#####end of program######

B. Splus/R code for Multivaraite Poisson Hidden Markov Model- Restricted and Independent Covariance Structure (This is an example for three components)

####Markov-dependent+ 3 components+ restricted###
##**NOTE: Before you start implementing the code please read the Chapter 5**##

```
data<-read.table("j://data1.txt",header=T) # Read the data from a text file
 attach(data)
 y1<-data[,1]
 y2<-data[,2]
 y3<-data[,3]
```

##Function to calculate bivariate Poisson probabilities
#[**Refer to Section 5.2.1 and equation (5.4)**]

$$p(\mathbf{y};\boldsymbol{\theta}) = P[Y_1 = y_1, Y_2 = y_2, ..., Y_n = y_n]$$

$$= \exp\left(-\sum_{i=1}^{n}\theta_i\right)\prod_{i=1}^{n}\frac{\theta_i^{y_i}}{y_i!}\sum_{i=0}^{s}\left[\prod_{j=1}^{n}\binom{y_j}{i} i!\left(\frac{\theta_0}{\prod_{k=1}^{n}\theta_k}\right)^i\right], \qquad (5.4)$$

where $s = \min\{y_1, y_2, ..., y_n\}$.

```
bivpois<-function(g1, g2, thetapar = c(1, 1, 1))
# calculates the probability function of a bivariate Poisson distribution
#with parameters thetapar = (theta1, theta2, theta3). The arguments g1 and g2 are
#the values of the two variables
{
# g1,g2 the two variables
      n <- length(g1)
      maxs <- c(max(g1), max(g2))       #Set initial values for parameters
         mins<-min(g1,g2)
      theta1 <- thetapar[1]
      theta2 <- thetapar[2]
      theta3 <- thetapar[3]
      thetasum<-sum(thetapar)
      prob <- matrix(NA, nrow = maxs[1] + 1, ncol = maxs[2] + 1, byrow =T)
      prob[1,1]<-exp( - thetasum)
      if((g1 == 0) | (g2 == 0)) {
            prob <- matrix(NA, nrow = maxs[1] + 1, ncol = maxs[2] + 1, byrow = T)
```

```
        prob[g1+1, g2 + 1] <- exp( - theta3) * dpois(g1, theta1) * dpois(g2,theta2)
    }
    else
    {
    k <- 1
    m <- 1
        for(i in 2:(maxs[1] + 1)) {
            prob[i, 1] <- (prob[i - 1, 1] * theta1)/(i - 1)
        }
        for(j in 2:(maxs[2] + 1)) {
            prob[1, j] <- (prob[1, j - 1] * theta2)/(j - 1)
        }
        for(j in 2:(maxs[2] + 1)) {
            for(i in 2:(maxs[1] + 1)) {
                prob[i, j] <- (theta1 * prob[i - 1, j] +
                  theta3 * prob[i - 1, j - 1])/(i - 1)
            }
        }
    }
    result <- prob
    result
}
```

##end of bivariate probability calculation

##Function to calculate trivariate Poisson probabilities
#[**Refer to Section 5.2.2**]

The joint probability function is given by

$$p(y_1,y_2,y_3;\boldsymbol{\theta}) = P[Y_1=y_1,Y_2=y_2,Y_3=y_3] = \sum_{\mathbf{x}^{(3)}}^{\mathbf{L}^{(2)}} \exp(-\sum_{m\in\mathbf{A}}\theta_m) \frac{\prod_{j\in R_1}\theta_j^{(y_j-\sum_{k\in R_2^{(j)}}x_k)} \prod_{i\in R_2}\theta_i^{x_i}}{\prod_{j\in R_1}(y_j-\sum_{k\in R_2^{(j)}}x_k)!\prod_{i\in R_2}x_i!} \quad .$$

```
threepois<-function(g1, g2, g3, thetapar = c(1, 1, 1, 1, 1, 1))
{
# calculates the probability function of a 3-variate Poisson distribution
#with parameters thetapar = (theta1, theta2, theta3,theta12,theta13,theta23). The
#arguments g1,  g2, g3 are the values of the two variables
    maxs <- c(max(g1), max(g2),max(g3))       #Set initial values for parameters
    mins<-min(g1,g2,g3)
```

```
      theta1 <- thetapar[1]
      theta2 <- thetapar[2]
      theta3 <- thetapar[3]
      theta12<-thetapar[4]
      theta13<-thetapar[5]
      theta23<-thetapar[6]
      thetasum<-sum(thetapar)
         prob <- array(0, dim=c(maxs[1]+1,maxs[2]+1,maxs[3]+1))
         tempor<-matrix(0,max(g1,g2,g3)+1,max(g1,g2,g3)+1)
         prob[1,1,1]<-exp( - thetasum)
tempor<-bivpois( maxs[2]+1, maxs[3]+1, c(theta2, theta3, theta23))
            for (k in 1:(maxs[3] + 1)) {
                  for (j in 1:(maxs[2] + 1)) {
prob[1, j, k]<-exp(-theta12-theta13)*dpois(0, theta1)*tempor[j,k]}
}
tempor<-bivpois( maxs[1]+1, maxs[2]+1, c(theta1, theta2, theta12))
#[Refer to Recurrence relationship equation (5.7)]
           for (i in 1:(maxs[1] + 1)) {
                  for (j in 1:(maxs[2] + 1)) {
prob[i,j,1]<-exp(-theta23-theta13)*dpois(0,theta3)*tempor[i,j]}}
tempor<-bivpois(maxs[1]+1, maxs[3]+1, c(theta1, theta3, theta13))
           for (i in 1:(maxs[1] + 1)) {
                  for (k in 1:(maxs[3] + 1)) {
prob[i,1,k]<-exp(-theta12-theta23)*dpois(0,theta2)*tempor[i, k]}}
            for (k in 1:(maxs[3] + 1)) {
                  for (j in 1:(maxs[2] + 1)) {
                     for (i in 1:(maxs[1] + 1)) {
                     if ((i-1)>0)         prob[i,j,k]<-prob[i-1,j,k]*theta1
                     if (((i-1)>0)&((j-1)>0))   prob[i,j,k]<-prob[i,j,k]+prob[i-1, j-
1,k]*theta12
                     if (((i-1)>0)&((k-1)>0))   prob[i,j,k]<-prob[i,j,k]+prob[i-1, j,k-
1]*theta13
                          if ((i-1)>0)  prob[i,j,k]<-prob[i,j,k]/(i-1)            }
           }
      }
      result <- prob
      result
}
##end of trivariate probability function

# initialize parameter values for component 1, 2, and 3. For the independent covariance
#model initial parameters for the covariance terms assign to zero.
theta11 <-1
theta21 <-2
theta31 <-1
theta121<-0
```

```
theta131<-0
theta231<-0

theta12 <-1.5
theta22 <-3
theta32 <-2
theta122<-0.1
theta132<-0
theta232<-0

theta13 <-0.5
theta23 <-2.5
theta33 <-1.2
theta123<-0
theta133<-0
theta233<-0

#constants#
N=3       # Number of states
T=150   #Number of observations
Nit=200 # number of  Iterations

#Initial Transition matrix##
TRANS=matrix(c(0.9,0.05,0.05,0.1,0.8,0.1,0.75,0.05,0.2),nrow=N,ncol=N,byrow=T)

loglike<-rep(0,Nit)

#Main loop of the EM algorithm
for (nit in 1:Nit){

#Initialize matrices to store the probabilities in different stages
result1<-matrix(NA,nrow=T,ncol=1)
d131<-matrix(NA,nrow=T,ncol=1)
d131old<-matrix(NA,nrow=T,ncol=1)
x131<-matrix(NA,nrow=T,ncol=1)
d132<-matrix(NA,nrow=T,ncol=1)
d132old<-matrix(NA,nrow=T,ncol=1)
x132<-matrix(NA,nrow=T,ncol=1)
d133<-matrix(NA,nrow=T,ncol=1)
d133old<-matrix(NA,nrow=T,ncol=1)
x133<-matrix(NA,nrow=T,ncol=1)

result2<-matrix(NA,nrow=T,ncol=1)
d121<-matrix(NA,nrow=T,ncol=1)
d121old<-matrix(NA,nrow=T,ncol=1)
x121<-matrix(NA,nrow=T,ncol=1)
```

```
d122<-matrix(NA,nrow=T,ncol=1)
d122old<-matrix(NA,nrow=T,ncol=1)
x122<-matrix(NA,nrow=T,ncol=1)
d123<-matrix(NA,nrow=T,ncol=1)
d123old<-matrix(NA,nrow=T,ncol=1)
x123<-matrix(NA,nrow=T,ncol=1)

result3<-matrix(NA,nrow=T,ncol=1)
d231<-matrix(NA,nrow=T,ncol=1)
d231old<-matrix(NA,nrow=T,ncol=1)
x231<-matrix(NA,nrow=T,ncol=1)
d232<-matrix(NA,nrow=T,ncol=1)
d232old<-matrix(NA,nrow=T,ncol=1)
x232<-matrix(NA,nrow=T,ncol=1)
d233<-matrix(NA,nrow=T,ncol=1)
d233old<-matrix(NA,nrow=T,ncol=1)
x233<-matrix(NA,nrow=T,ncol=1)

#Initialize matrices to store probabilities from three states
threep11<-matrix(NA,nrow=T,ncol=1)
threep22<-matrix(NA,nrow=T,ncol=1)
threep33<-matrix(NA,nrow=T,ncol=1)

# Initialize matrices to store values of X's
x11<-matrix(NA,nrow=T,ncol=1)
x21<-matrix(NA,nrow=T,ncol=1)
x31<-matrix(NA,nrow=T,ncol=1)
x12<-matrix(NA,nrow=T,ncol=1)
x22<-matrix(NA,nrow=T,ncol=1)
x32<-matrix(NA,nrow=T,ncol=1)
x13<-matrix(NA,nrow=T,ncol=1)
x23<-matrix(NA,nrow=T,ncol=1)
x33<-matrix(NA,nrow=T,ncol=1)

maxs<-c(max(y1),max(y2),max(y3))
threep1<-array(0,dim=c(maxs[1]+1,maxs[2]+1,maxs[3]+1))
threep2<-array(0,dim=c(maxs[1]+1,maxs[2]+1,maxs[3]+1))
threep3<-array(0,dim=c(maxs[1]+1,maxs[2]+1,maxs[3]+1))

# start EM algorithm
for (i in 1:T){
threep1<-(threepois(y1[i],y2[i],y3[i],
thetapar=c(theta11,theta21,theta31,theta121,theta131,theta231)))
threep11[i]<-threep1[y1[i]+1,y2[i]+1,y3[i]+1]
threep2<-(threepois(y1[i],y2[i],y3[i],
thetapar=c(theta12,theta22,theta32,theta122,theta132,theta232)))
```

```
threep22[i]<-threep2[y1[i]+1,y2[i]+1,y3[i]+1]
threep3<-(threepois(y1[i],y2[i],y3[i],
thetapar=c(theta13,theta23,theta33,theta123,theta133,theta233)))
threep33[i]<-threep3[y1[i]+1,y2[i]+1,y3[i]+1]

d131[i]<-0
d132[i]<-0
d133[i]<-0
result1[i]<-min(y1[i],y3[i])
for (r in 0:result1[i]){
d131old[i]<-r*dpois(y1[i]-r,theta11)*dpois(y3[i]-r,theta31)*dpois(r,theta131)
d131[i]<-d131old[i]+d131[i]
d132old[i]<-r*dpois(y1[i]-r,theta12)*dpois(y3[i]-r,theta32)*dpois(r,theta132)
d132[i]<-d132old[i]+d132[i]
d133old[i]<-r*dpois(y1[i]-r,theta13)*dpois(y3[i]-r,theta33)*dpois(r,theta133)
d133[i]<-d133old[i]+d133[i]}


d121[i]<-0
d122[i]<-0
d123[i]<-0
result2[i]<-min(y1[i],y2[i])
for (r in 0:result2[i]){
d121old[i]<-r*dpois(y1[i]-r,theta11)*dpois(y2[i]-r,theta21)*dpois(r,theta121)
d121[i]<-d121old[i]+d121[i]
d122old[i]<-r*dpois(y1[i]-r,theta12)*dpois(y2[i]-r,theta22)*dpois(r,theta122)
d122[i]<-d122old[i]+d122[i]
d123old[i]<-r*dpois(y1[i]-r,theta13)*dpois(y2[i]-r,theta23)*dpois(r,theta123)
d123[i]<-d123old[i]+d123[i]}


d231[i]<-0
d232[i]<-0
d233[i]<-0
result3[i]<-min(y2[i],y3[i])
for (r in 0:result3[i]){
d231old[i]<-r*dpois(y2[i]-r,theta21)*dpois(y3[i]-r,theta31)*dpois(r,theta231)
d231[i]<-d231old[i]+d231[i]
d232old[i]<-r*dpois(y2[i]-r,theta22)*dpois(y3[i]-r,theta32)*dpois(r,theta232)
d232[i]<-d232old[i]+d232[i]
d233old[i]<-r*dpois(y2[i]-r,theta23)*dpois(y3[i]-r,theta33)*dpois(r,theta233)
d233[i]<-d233old[i]+d233[i]}

# Check the condition for Poisson random variables
#[Refer to equation (5.2)]
```

$$
\begin{aligned}
&y_1 - x_{12} - x_{13} - x_{123} \geq 0 \\
&y_2 - x_{12} - x_{23} - x_{123} \geq 0 \\
&y_3 - x_{13} - x_{23} - x_{123} \geq 0.
\end{aligned} \tag{5.2}
$$

if (threep11[i]==0) {x131[i]<-0} else {x131[i]<-d131[i]/threep11[i]}

if (threep11[i]==0) {x121[i]<-0} else {x121[i]<-d121[i]/threep11[i]}

if (threep11[i]==0) {x231[i]<-0} else {x231[i]<-d231[i]/threep11[i]}

if (threep22[i]==0) {x132[i]<-0} else {x132[i]<-d132[i]/threep22[i]}

if (threep22[i]==0) {x122[i]<-0} else {x122[i]<-d122[i]/threep22[i]}

if (threep22[i]==0) {x232[i]<-0} else {x232[i]<-d232[i]/threep22[i]}

if (threep33[i]==0) {x133[i]<-0} else {x133[i]<-d133[i]/threep33[i]}

if (threep33[i]==0) {x123[i]<-0} else {x123[i]<-d123[i]/threep33[i]}

if (threep33[i]==0) {x233[i]<-0} else {x233[i]<-d233[i]/threep33[i]}

if ((y1[i]-x131[i]-x121[i])>0) {x11[i]<-(y1[i]-x131[i]-x121[i])} else {x11[i]<-y1[i]}

if ((y2[i]-x121[i]-x231[i])>0) {x21[i]<-(y2[i]-x121[i]-x231[i])} else {x21[i]<-y1[i]}

if ((y3[i]-x131[i]-x231[i])>0) {x31[i]<-(y3[i]-x131[i]-x231[i])} else {x31[i]<-y1[i]}

if ((y1[i]-x132[i]-x122[i])>0) {x12[i]<-(y1[i]-x132[i]-x122[i])} else {x12[i]<-y2[i]}

if ((y2[i]-x122[i]-x232[i])>0) {x22[i]<-(y2[i]-x122[i]-x232[i])} else {x22[i]<-y2[i]}

if ((y3[i]-x132[i]-x232[i])>0) {x32[i]<-(y3[i]-x132[i]-x232[i])} else {x32[i]<-y2[i]}

if ((y1[i]-x133[i]-x123[i])>0) {x13[i]<-(y1[i]-x133[i]-x123[i])} else {x13[i]<-y3[i]}

if ((y2[i]-x123[i]-x233[i])>0) {x23[i]<-(y2[i]-x123[i]-x233[i])} else {x23[i]<-y3[i]}

if ((y3[i]-x133[i]-x233[i])>0) {x33[i]<-(y3[i]-x133[i]-x233[i])} else {x33[i]<-y3[i]}}

#Initialize forward and backward variables and place to store loglikelihood#
logl=matrix(0,nrow=1,ncol=Nit)

```
dens=matrix(0,nrow=T,ncol=N)
alpha=matrix(0,nrow=T,ncol=N)
beta=matrix(0,nrow=T,ncol=N)
scale=matrix(c(1,rep(0,T-1)),nrow=1,ncol=T)
ones=matrix(1,nrow=T,ncol=1)

####E-step, compute density values
        dens=cbind(threep11,threep22,threep33)

####E-step, forward recursion and likelihood computation
##Use a uniform a priori probability for the initial state
#[Refer to equation (5.25) and (5.26)]
#The forward and backward variables
```

$\# \alpha_j(i) = P[\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_n, S_i = j]$ and

$$\# \beta_j(i) = P[\mathbf{y}_{i+1}, ..., \mathbf{y}_n \mid S_i = j] \tag{5.25}$$

#which yield the quantities of interest by

$$\# \hat{u}_j(i) = \frac{\alpha_j(i)\beta_j(i)}{\sum_l \alpha_l(n)} = \frac{\alpha_j(i)\beta_j(i)}{\sum_{j=1}^{m} \alpha_j(i)\beta_j(i)} \quad \text{and}$$

$$\# \hat{v}_{jk}(i) = \frac{P_{jk} f(\mathbf{y}_i; \boldsymbol{\lambda}_k)\alpha_j(i-1)\beta_k(i)}{\sum_l \alpha_l(n)}. \tag{5.26}$$

```
        alpha[1,]=dens[1,]/N
        for (t in 2:T){
                alpha[t,]=(alpha[t-1,]%*%TRANS)*dens[t,]
        #####Systematic scaling
                scale[,t]=sum(alpha[t,])
                alpha[t,]=alpha[t,]/scale[,t]
        }

###compute likelihood
loglike[nit]<-sum(log(scale))

####E-step, backward recursion
###Scale the backward variable with the forward scale factors
###(this ensures that the reestmation of the transition matrix below is correct)
#[Refer to equation (5.25) and (5.26)]
```

```
beta[T,]=matrix(1,nrow=1,ncol=N)
for (t in (T-1):1){
        beta[t,]=(beta[t+1,]*dens[t+1,])%*%t(TRANS)
        beta[t,]=beta[t,]/scale[,t]
}
```

####M-step, reestmation of the transition matrix
##compute unnormalized transition probabilities (this is indeed still the end of the E-
#step, which explains that TRANS appears on the right-hand side below)
#[**Refer to equation (5.24)**]

$$P_{jk} = \frac{\sum_{i=2}^{n} \hat{v}_{jk}(i)}{\sum_{i=2}^{n}\sum_{l=1}^{m} \hat{v}_{jl}(i)} \ . \qquad (5.24)$$

```
TRANS=TRANS*(t(alpha[1:(T-1),])%*%(dens[2:T,]*beta[2:T,]))
```

###Normalization of the transition matrix

```
oness=matrix(1,nrow=1,ncol=N)
sumtrans=matrix(0,nrow=N,ncol=1,byrow=T)
for (n in 1:N){
sumtrans[n,]=sum(TRANS[n,])
}
sumtrans=sumtrans%*%oness
TRANS=TRANS/sumtrans
```

####M-step, reestimation of the rates
###Compute a posteriori probabilities and store them in matrix beta to save space
#[**Refer to equation (5.25) and (5.26)**]
#The forward and backward variables

$\alpha_j(i) = P[\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_n, S_i = j]$ and

$\beta_j(i) = P[\mathbf{y}_{i+1}, ..., \mathbf{y}_n \mid S_i = j]$ (5.25)

#which yield the quantities of interest by

$\hat{u}_j(i) = \frac{\alpha_j(i)\beta_j(i)}{\sum_l \alpha_l(n)} = \frac{\alpha_j(i)\beta_j(i)}{\sum_{j=1}^{m} \alpha_j(i)\beta_j(i)}$ and

$$\#\hat{v}_{jk}(i) = \frac{P_{jk} f(\mathbf{y}_i; \boldsymbol{\lambda}_k) \alpha_j(i-1) \beta_k(i)}{\sum_l \alpha_l(n)}. \quad (5.26)$$

```
beta=alpha*beta
sumbeta=matrix(0,nrow=T,ncol=1,byrow=T)
for (r in 1:T){
sumbeta[r,]=sum(beta[r,])
}
sumbeta=sumbeta%*%oness
beta=beta/sumbeta

##Reestmate rates
#component 1
newdata1=cbind(x11,x21,x31,x131,x121,x231)
rate1=(t(beta[,1])%*%newdata1)/sum(beta[,1])
#component 2
newdata2=cbind(x12,x22,x32,x132,x122,x232)
rate2=(t(beta[,2])%*%newdata2)/sum(beta[,2])
#component 3
newdata3=cbind(x13,x23,x33,x133,x123,x233)
rate3=(t(beta[,3])%*%newdata3)/sum(beta[,3])


#Assign estimated parameters to new variables
theta11 <-rate1[,1] #component 1 estimates
theta21 <-rate1[,2]
theta31 <-rate1[,3]
theta121<-rate1[,4]
theta131<-rate1[,5]
theta231<-rate1[,6]

theta12 <-rate2[,1] #component 2 estimates
theta22 <-rate2[,2]
theta32 <-rate2[,3]
theta122<-rate2[,4]
theta132<-rate2[,5]
theta232<-rate2[,6]

theta13 <-rate3[,1] #component 3 estimates
theta23 <-rate3[,2]
theta33 <-rate3[,3]
theta123<-rate3[,4]
theta133<-rate3[,5]
theta233<-rate3[,6]
}
```

####RESULTS####
loglike
rate1
rate2
rate3
TRANS
######end of program########

For any further questions:
Email:cpk646@mail.usask.ca